Hu *et al. Comput Soc Netw* (2019) 6:10
https://doi.org/10.1186/s40649-019-0071-4

Computational Social Networks

RESEARCH Open Access

An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning

Han Hu[1], NhatHai Phan[1*], Soon A. Chun[2], James Geller[1], Huy Vo[3], Xinyue Ye[1], Ruoming Jin[4], Kele Ding[4], Deric Kenne[4] and Dejing Dou[5]

*Correspondence: phan@njit.edu
[1] New Jersey Institute of Technology, University Heights, Newark 07102, USA
Full list of author information is available at the end of the article

Abstract

Drug abuse continues to accelerate towards becoming the most severe public health problem in the United States. The ability to detect drug-abuse risk behavior at a population scale, such as among the population of Twitter users, can help us to monitor the trend of drug-abuse incidents. Unfortunately, traditional methods do not effectively detect drug-abuse risk behavior, given tweets. This is because: (1) tweets usually are noisy and sparse and (2) the availability of labeled data is limited. To address these challenging problems, we propose a deep self-taught learning system to detect and monitor drug-abuse risk behaviors in the Twitter sphere, by leveraging a large amount of unlabeled data. Our models automatically augment annotated data: (i) to improve the classification performance and (ii) to capture the evolving picture of drug abuse on online social media. Our extensive experiments have been conducted on three million drug-abuse-related tweets with geo-location information. Results show that our approach is highly effective in detecting drug-abuse risk behaviors.

Keywords: Deep learning, Self-taught learning, Drug abuse, Twitter

Introduction

Abuse of prescription drugs and of illicit drugs has been declared a "national emergency" [1]. This crisis includes the misuse and abuse of cannabinoids, opioids, tranquilizers, stimulants, inhalants, and other types of psychoactive drugs, which statistical analysis documents as a rising trend in the United States. The most recent reports from the National Survey on Drug Use and Health (NSDUH) [2] estimate that 10.6% of the total population of people ages 12 years and older (i.e., about 28.6 million people) misused illicit drugs in 2016, which represents an increase of 0.5% since 2015 [3]. According to the Centers for Disease Control and Prevention (CDC), opioid drugs were involved in 42,249 known deaths in 2016 nationwide [4]. In addition, the number of heroin-involved deaths has been increasing sharply for 5 years, and surpassed the number of firearm homicides in 2015 [5].

In April 2017, the Department of Health and Human Services announced their "Opioid Strategy" to battle the country's drug-abuse crisis [1]. In the Opioid Strategy, one of the major aims is to strengthen public health data collection, to inform a timeliness

public health response, as the epidemic evolves. Given its 100 million daily active users and 500 million daily *tweets* [6] (messages posted by Twitter users), Twitter has been used as a sufficient and reliable data source for many detection tasks, including epidemiology [7] and public health [8–13], at the population scale, in a real-time manner. Motivated by these facts and the urgent needs, our goal in this paper is to develop a large-scale computational system to detect drug-abuse risk behaviors via Twitter sphere.

Several studies [10, 13–17] have explored the detection of prescription drug abuse on Twitter. However, the current state-of-the-art approaches and systems are limited in terms of scales and accuracy. They typically applied keyword-based approaches to collect tweets explicitly mentioning specific drug names, such as Adderall, Oxycodone, Quetiapine, Metformin, Cocaine, marijuana, weed, meth, tranquilizer, etc. [10, 13, 15, 17]. However, that may not reflect the actual distribution of drug-abuse risk behaviors on online social media, since: (1) the expressions of drug-abuse risk behaviors are often vague, in comparison with common topics, i.e., a lot of slang is used and (2) relying on only keyword-based approaches is susceptible to lexical ambiguity in natural language [12]. In addition, the drug-abuse risk behavior Twitter data are very imbalanced, i.e., dominated by non-drug-abuse risk behavior tweets, such as drug-related news, social discussions, reports, advertisements, etc. The limited availability of annotated tweets makes it even more challenging to distinguish drug-abuse risk behaviors from drug-related tweets. However, existing approaches [10, 13, 15, 17] have not been designed to address these challenging issues for drug-abuse risk behavior detection on online social media.

Contributions: To address these challenges, our main contributions are to propose: (1) a large-scale drug-abuse risk behavior tweets collection mechanism based on supervised machine-learning and data crowd-sourcing techniques and (2) a deep self-taught learning algorithm for drug-abuse risk behavior detection. Based on our previous work [18], we extended the analysis of the classification results from our three million tweets dataset with the analysis of word frequency, hashtag frequency, drug name-behavior co-occurrence, temporal distribution (time-of-day), and state-level spatial distribution.

We first collect tweets through a filter, in which a variety of drug names, colloquialisms and slang terms, and abuse-indicating terms (e.g., *overdose, addiction, high, abuse*, and even *death*) are combined together. We manually annotate a small number of tweets as seed tweets, which are used to train machine-learning classifiers. Then, the classifiers are applied to large number of unlabeled tweets to produce machine-labeled tweets. The machine-labeled tweets are verified again by humans on Mechanical Turk, i.e., a crowd-sourcing platform, with good accuracy but at a much lower cost. The new labeled tweets and the seed tweets are combined to form a sufficient and reliable labeled dataset for drug-abuse risk behavior detection by applying deep learning models, i.e., convolution neural networks (CNN) [19], long–short-term memory (LSTM) models [20], etc.

However, there are still a large amount of unlabeled data, which can be leveraged to significantly improve our models in terms of classification accuracy. Therefore, we further propose a self-taught learning algorithm, in which the training data of our deep self-taught learning models will be recursively augmented with a set of new machine-labeled tweets. These machine-labeled tweets are generated by applying the previously trained

deep learning models to a random sample of a huge number of unlabeled tweets, i.e., the three million tweets in our dataset. Note that the set of new machine-labeled tweets possibly has a different distribution from the original training and test datasets.

After the model is trained, we apply it to our geo-location-tagged dataset to acquire classification results for analysis. Results from the aforementioned analysis show that the drug-abuse risk behavior-positive tweets have distinctive patterns of words, hashtags, drug name-behavior co-occurrence, time-of-day distribution, and spatial distribution, compared with other tweets. These results show that our approach is highly effective in detecting drug-abuse risk behaviors.

The rest of this paper is organized as follows. We review related work in the next section. Then, we describe the implementation of our method in detail, followed by experiment results and data analysis. Finally, we conclude this paper and propose future directions.

Background and related work

On one hand, the traditional studies and organizations, such as NSDUH [2], CDC [4], Monitoring the Future [21], the Drug-Abuse Warning Network (DAWN) [22], and the MedWatch program [23] are trustworthy sources for getting the general picture of the drug-abuse epidemic. On the other hand, many studies that are based on modern online social media, such as Twitter, have shown promising results in drug-abuse detection and related topics [7–17]. Many of the existing studies were focusing on the quantitative analysis utilizing data from online social media. Meng et al. [24] used traditional text and sentiment analysis methods to investigate substance use patterns and underage use of substance, and the association between demographic data and these patterns. Ding et al. [25] investigated the correlation between substance (tobacco, alcohol, and drug) use disorders and words in Facebook users' "Status Updates" and "Likes". Their results showing word patterns are different between users who have substance us disorder and users who do not have. Hanson et al. [15] conducted a quantitative analysis on 213,633 tweets discussing "Adderall", a prescription stimulant commonly abused among college students, and also published another study [14] focused on how possible drug-abuser interact with and influence others in online social circles. The results showed that strong correlation could be found: (1) between the amount of interaction about prescription drugs and a level of abusiveness shown by the network and (2) between the types of drugs mentioned by the index user and his or her network. Shutler et al. [17] performed a qualitative analysis of six prescription opioids, i.e., Percocet, Percs, OxyContin, Oxys, Vicodin, and Hydros. Tweets were collected with exact word matching and manual classification. Their primary goal was to identify the key terms used in tweets that likely indicate drug abuse. They found that the use of Oxys, Percs, and OxyContin was common among the tweets, where there were positive indications of abuse. McNaughton et al. [16] measured online endorsement of prescription opioid abuse by developing an integrative metric through the lens of Internet communities. Simpson et al. [26] demonstrated an attempt to identify emerging drug terms using NLP techniques. Furthermore, Twitter and social media have been shown to be reliable sources in analyzing drug abuse and public health-related topics, such as cigarette smoking [8, 12], alcohol use [11], and even cardiac arrest [9]. However, these studies generally did not propose methods that apply to large-scale automated monitoring tasks.

Our previous work [27] showed the potential of applying machine-learning models in a drug-abuse monitoring system to detect drug-abuse-related tweets. Several other approaches also utilized machine-learning methods in detecting and analyzing drug-related posts on Twitter. For instance, Coloma et al. [28] illustrated the potential of social media in drug safety surveillance with two case study multiple online social media platforms. Sarker et al. [13] proposed a supervised classification model, in which different features such as n-grams, abuse-indicating terms, slang terms, synonyms, etc., were extracted from manually annotated tweets. Then, these features were used to train traditional machine-learning models to classify drug-abuse tweets and non-abuse tweets. Recently, many works, including one of our works [29], explored the use of more advanced deep learning models for drug-related classification tasks on online social media. Following our work, Kong et al. [30] proposed deep learning model that utilizes geographical prior information as input features. Chary et al. [10] discussed how to use AI models to extract content useful for purposes of toxicovigilance from social media, such as Facebook, Twitter, and Google+. Weissenbacher et al. [31] proposed deep neural network based model to detect drug name mentions in tweets. Mahata et al. [32] proposed an ensemble CNN model to classify tweets from three classes, i.e., personal medication intakes, possible personal medication intake, and non-intake. Works have also been done in perspectives other than content-based analysis and classification. Zhang et al. [33] proposed a complex schema, which models all possible interactions between users and posts, for automatic detection of drug abusers on Twitter. Li et al. [34] evaluated deep learning models against traditional machine-learning models on the task of detecting illicit drug dealers on Instagram.

Although existing studies have shown promising approaches towards the detecting of drug-related posts and information on popular online social media platforms, such as Twitter and Instagram, their limitations can be identified as: (1) limited in scale, as the methods proposed in many studies do not scale well, or rely on larger manually annotated training dataset for higher performance; (2) limited in scope, as most studies focus on a small group of drugs; and (3) limited in performance, as many methods use traditional machine-learning models. In this paper, we propose a novel deep self-taught learning system to leverage a huge number of unlabeled tweets. Self-taught learning [35] is a method that integrated the concepts of semi-supervised and multi-task learning, in which the model can exploit examples that are unlabeled and possibly come from a distribution different from the target distribution. It has already been shown that deep neural networks can take advantage of unsupervised learning and unlabeled examples [36, 37]. Different from other approaches mainly designed for image processing and object detection [38–41], our deep self-learning model shows the ability to detect drug-abuse risk behavior given noisy and sparse Twitter data with a limited availability of annotated tweets.

Deep self-taught learning system for drug-abuse risk behavior detection

In this section, we present the definition of the drug-abuse risk behavior detection problem, our system for collecting tweets, labeling tweets, and our deep self-taught learning approach. The system overview is shown in Fig. 1.

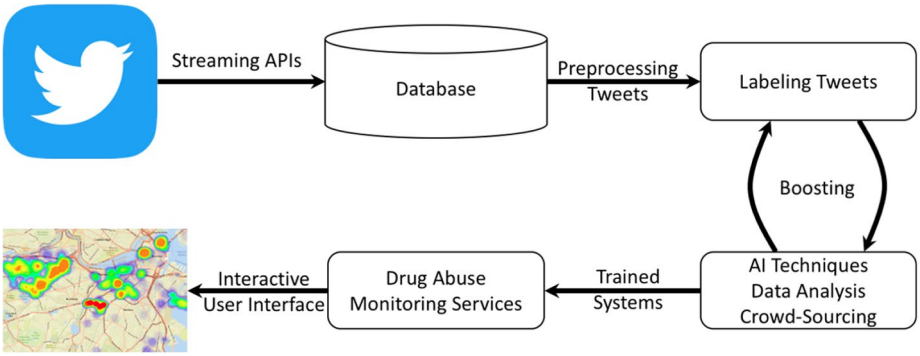

Fig. 1 Drug-abuse detection system. There are 4 steps as follows: (1) Tweets will be collected through Twitter APIs. (2) Pre-processed tweets will be labeled by humans, AI techniques, and crowd-sourcing techniques. (3) Labeled tweets will be used to augment the training data of our AI models and data analysis tasks to identify tweets with drug-abuse risk behaviors, through a self-taught algorithm. In addition, (4) trained systems will be used in different drug-abuse monitoring services and interactive user interfaces

Problem definition

We use the term "drug-abuse risk behavior" in the wider sense, including misuse and use of Schedule 1 drugs that are illegal; and misuse of Schedule 2 drugs, e.g., *Oxycodone*, which includes the use thereof for non-medical purposes, and the symptoms and side-effects of misuse. Our task is to develop classification models that can classify a given unlabeled tweet into one of the two classes: a drug-abuse risk behavior tweet (*positive*) or a non-drug-abuse risk behavior (*negative*) tweet. The main criteria for classifying a tweet as drug-abuse risk can be condensed into: "*The existence of abusive activities or endorsements of drugs.*" Meanwhile, news, reports, and opinions about drug abuse are the signals of tweets that are not considered as containing abuse risk.

Collecting and labeling tweets

In our crawling system, raw tweets are collected through Twitter APIs. For the collection of focused Twitter data, we use a list of the names of illegal and prescription drugs [42] that have been commonly abused over time, e.g., *Barbiturates, OxyContin, Ritalin, Cocaine, LSD, Opiates, Heroin, Codeine, Fentanyl*, etc. However, the data are very noisy, since: (1) there is no indication of how to distinguish between drug abuse and legitimate use (of prescription drugs) in collected Tweets and (2) many of slang terms are used in expressing drug-abuse risk behavior. To address this problem, we added slang terms for drugs and abuse-indicating terms, e.g., "high," "stoned," "blunt," "addicted," etc., into our keyword search library. These slang terms are clearly expressing that the tweets in question were about drug abuse. As a result, most of the collected data are drug abuse-related.

To obtain trustworthy annotated data, we design two integrative steps in labeling tweets. In the first step, 1,794 tweets randomly chosen from collected tweets were manually annotated as positive or negative by three team members who have experience in health informatics. Several instances of positive tweets and negative tweets are illustrated in Table 1. These labeled tweets are considered seed tweets, which then are used

Table 1 Instances of manually annotated positive tweets and negative tweets

	Tweets
Positive	"Ever since my Acid trips like whenever I get super high I just start—lightly hallucinating and it's tbh creepy"
	"drove like 10 miles on these icy ass roads all to get some weed if imma—be locked up in my house for awhile imma need some weed"
	"Smoking a blunt at home so much better than going to the woods in—Brooksville and puking on yourself Bc you drank too much reball"
Negative	"Just watched Fear and Loathing in Las Vegas for the first time—and I think I should have been on acid to fully understand it"
	"today I was asked if I do heroin because I went to Lancaster????"
	"Morgan told me my Bitmoji looks like a heroin addict?"

to train traditional binary classifiers, e.g., SVM, Naive Bayes, etc., to predict whether a tweet is a drug-abuse risk behavior tweet or not. The trained classifiers are applied to unlabeled tweets to predict their labels, which are called machine labels. In the second step, 5000 positive machine-labeled tweets with high classification confidence are verified again on Amazon Mechanical Turk (AMT), which is a well-known crowd-sourcing platform. To improve the trustworthiness and to avoid bias in the annotated data, each tweet is labeled by three individual workers. The workers are instructed to follow with the same annotation instructions that our annotators have followed. Our annotators also labeled a random sample of 1000 tweets and compare the labels with the results from AMT, as a quality check.

Tweet vectorization

Raw tweets need to be first pre-processed, then represented as vectors, before they can be used in training machine-learning models. In this study, we choose a commonly used pre-processing pipeline, followed by three different vectorization methods. The pre-processing pipeline consists of following steps:

- The tweets are tokenized and lower cased. The special entities, i.e., including Emojis, URLs, mentions, and hashtags, are removed or replaced with special keywords. The non-word characters, i.e., including HTML symbols, punctuation marks, and foreign characters, are removed. Words with three or more repeating characters are reduced to at most three successive characters.
- Stop words are removed according to a custom stop-word list. Stemming is applied using the standard Porter Stemmer.

After the pre-processing steps, common vectorization methods are used to extract features from tweets, including: (1) term frequency, denoted as *tf*, (2) *Tf-idf*, and (3) Word2vec [43]. Word2vec is an advanced and effective word embedding method that converts each word into a dense vector of fixed length. We considered two different word2vec models: (i) a custom word2vec model, which was trained on our three million drug-abuse-related tweets. The model contains 300-dimensional vectors for 1,130,962 words and phrases and (ii) Google word2vec, which is a well-known

pre-trained word2vec model built from part of a Google News dataset with about 100 billion words, and the model contains 300-dimensional vectors for three million words and phrases.

Deep self-taught learning approach

By applying both traditional and advanced machine-learning models, such as SVM, Naive Bayes, CNN, and LSTM to the small and static annotated data, i.e., 6794 tweets, we can achieve reasonable classification accuracies of nearly 80%, as indicated in Fig. 3 when the number of iteration k is zero, which is equivalent to applying models without the proposed self-taught method. However, to develop a scalable and trustworthy drug-abuse risk behavior detection model, we need to: (1) improve classification models to achieve higher accuracy and performance and (2) leverage the large number of unlabeled tweets, i.e., three million tweets related to drug abuse, to improve the system performance. Therefore, we propose a deep self-taught learning model by repeatedly augmenting the training data with machine-labeled tweets. The pseudo-code of our algorithm is as follows:

Step 1: Randomly initialize labeled data D consisting of 5794 annotated tweets as the training set. Initialize a test data T consisting of the remaining 1000 annotated tweets.

Step 2: Train a binary classification model M using the labeled data D. M could be a CNN model or an LSTM model.

Step 3: Use the model M to label the unlabeled data U, which simply consists of three million unlabeled tweets. The set of new labeled tweets is denoted as $\overline{D}$, which is also called machine-labeled data.

Step 4: Sample tweets from the machine-labeled data $\overline{D}$ with a high classification confidence, and then add the sampled tweets D^+ into the labeled data D to form a new training dataset: $D = D \bigcup D^+$. A tweet is considered to have a high classification confidence if it has a classification probability $p \in [0, 1]$ higher than a predefined sampling threshold δ. Sampled machine-labeled tweets will not be sampled again: $U = U - D^+$.

Step 5: Repeat Steps 2–4 for k iterations, where k is a user-predefined number. Return the trained model M.

With the self-taught learning method, the training data contain the annotated data D, which is automatically augmented with highly confident, machine-labeled tweets, in each iteration. This approach has the potential of increasing the classification performance of our model over time. In addition, the unlabeled data can be collected from the Twitter APIs in real time, to capture the evolving of English (slang) terms about drug-abuse risk behaviors. In the literature, data augmentation approaches have been applied to improve the accuracy of deep learning models [36]. However, the existing approaches [36, 39–41] are quite different from our proposed model, since they focused on image classification tasks, instead of drug-abuse risk

behavior detection as in our study. Note that, to ensure fairness, test data T are separated from other data sources during the training process.

Experimental results

To examine the effectiveness and efficiency of our proposed deep self-taught learning approaches, we have carried out a series of experiments using a set of three million drug-abuse-related tweets collected in the past 4 years. We first elaborate details about our dataset, baseline approaches, measures, and model configurations. Then, we introduce our experimental results.

Experiment settings

Dataset

The *seed dataset* contains 1794 tweets that were manually labeled by three annotators, including 280 positive tweets and 1514 negative tweets. The agreement score among three annotators is 0.414, measured by Krippendoff's Alpha. We then selected 5000 tweets labeled by the machine-learning model (i.e., SVM) with a high confidence level ($\delta > 0.7$), and rendered them verified on AMT. The AMT workers have the agreement score of 0.456, measured by Krippendoff's Alpha. Note that both agreement scores should be considered as reliable result in out study settings [44], since: (1) our task is to reduce variability in data annotation, instead of typical content analysis [45] and (2) the Krippendoff's Alpha is sensitive to data imbalance and sparseness, which are the characteristics of our dataset. Our integrative labeling approach resulted in a reliable and well-balanced annotated dataset, with 6794 labeled tweets, including 3102 positive labels and 3677 negative labels. For the unlabeled data, we have the three million drug-abuse-related tweets with geo-location information covering the entire continental US (lower 48 states and D.C.).

Baseline methods

In our experiments, Random Forest (*RF*), Naive Bayes (*NB*), and *SVM* are employed as baseline approaches for the binary classification task, i.e., to classify whether a tweet is a drug-abuse risk behavior tweet or not. Table 2 shows the parameter settings of baseline approaches and the proposed models. Note that for the Naive Bayes method, we use Gaussian Naive Bayes with word2vec embedding. Meanwhile, we use term frequency (i.e., *tf*) and *tf-idf* vectorization for Multinominal Naive Bayes. This is because: (1) the vectors generated by term frequency-based vectorization has a very high number of dimensions and could be only represented by sparse matrix, which was not supported by the chosen implementation of Gaussian Naive Bayes and (2) the multinominal Naive Bayes require non-negative inputs, but vectors generated by word2vec embedding have negative values. Regarding our self-taught CNN (*st-CNN*) and self-taught LSTM (*st-LSTM*) models, the Adam optimizer algorithm with default learning rate is used for training. The number of iterations k is set to 6, and the sampling threshold δ is set to 0.7, for all methods. All the experiments have been conducted on a single GPU, i.e., NVIDIA TITAN Xp with 12 GB memory and 3072 CUDA cores.

Table 2 Parameter settings for all models

Baseline model	Parameter setting	
SVM	$C = 5.0$, gamma = 0.01, kernel:rbf	
Random Forest	N_estimators = 500, class_weight = balanced, max_depth = 20	
Naive Bayes (Gaussian)	Default	
Naive Bayes (multinominal)	Default	
Proposed model	**Layers**	**Parameter setting**
Self-taught CNN (st-CNN)	Embedding	Size: 300, max_length: 20
	Dropout	Dropout_rate: 0.2
	Convolutional	Kernel_sizes: [2,3,4], number_kernels: 20 activation_function: Relu, strides: 1
	Max pooling	Pool_size: 2
	Flatten	No parameter
	Concatenate	No parameter
	Dropout	Dropout_rate: 0.5
	Two dense layers	Dense_layer_1: size: 520 × 500; dense_layer_2: size: 500 × 2
Self-taught LSTM (st-LSTM)	Embedding	Size: 300, max_length: 20
	Dropout	Dropout_rate: 0.2
	LSTM	Sequence_output: false
	Dropout	Dropout_rate: 0.5
	Two dense layers	Dense_layer_1: size: 300 × 500; dense_layer_2: size: 500 × 2

Measures

Accuracy, recall, and F1-value are used to validate the effectiveness of the proposed and baseline approaches. Due to the small size and the imbalanced label distribution, we adopted the Monte Carlo cross-validation technique. In each run, a fixed number of data instances are sampled (i.e., without replacement) as the test dataset, and the rest of the data as the training dataset. Multiple runs (i.e., 3 times) are generated for each model in each set of parameters and experimental configurations. We report the average of these runs as result. Definitions of the accuracy, recall, and F1-value are given as follows, where T_P,T_N,F_P,F_N are the number of true positives, true negatives, false positives, and false negatives, correspondingly.

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N};\ \text{Recall} = \frac{T_P}{T_P + F_N};\ \text{F1-value} = \frac{2T_P}{2T_P + F_P + F_N}$$

Experiment questions

Our task of validation concerns three key issues: (1) which parameter configurations are optimal for the baseline models on the seed dataset, i.e., SVM, RF, and NB? (2) which self-taught learning model is the best in terms of accuracy, recall, and F1-value, given the 6794 annotated tweets and the three million unlabeled tweets? and (3) which vectorization setting is more effective? To address these concerns, our series of experiments are as follows.

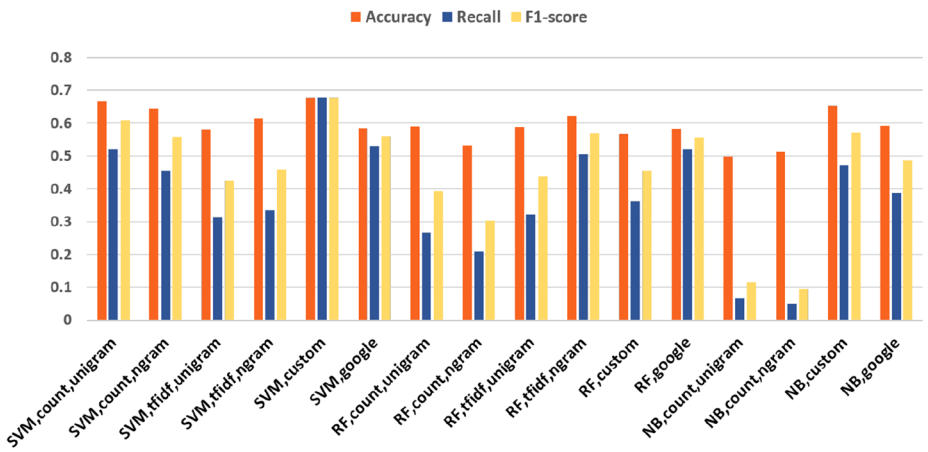

Fig. 2 Accuracy, recall, and F1-value of each baseline model on the seed dataset

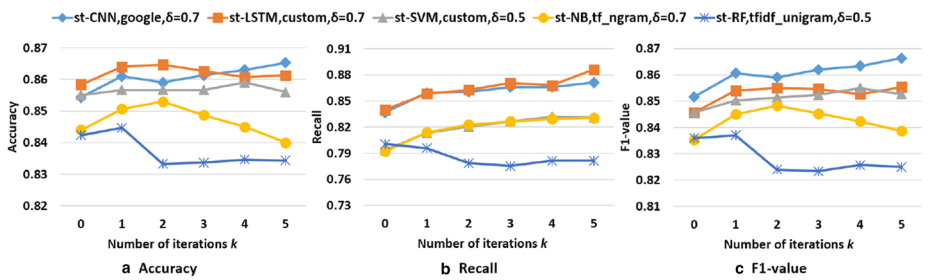

Fig. 3 Accuracy, recall, and F1-value of the five self-taught learning models, including st-CNN, st-LSTM, st-SVM, st-NB, and st-RF

Experimental results

Experiment on seed dataset with baseline models

Figure 2 illustrates the accuracy, recall, and F1-value of each algorithm with different parameter configurations, i.e., term frequency *tf*, *tf-idf*, and *word2vec*, on the (annotated) seed dataset. The term "*custom*" is used to indicate the word2vec embedding trained with our own drug-abuse-related tweets, compared with the pre-trained Google News word2vec embedding, denoted as "*google*." It is clear that the SVM model using the custom-trained word2vec embedding achieves the best and the most balanced performance in terms of all three measures, i.e., accuracy, recall, and F1-value, at approximately 67%. Other configurations usually have a lower recall, which suggests that the decisions they make bias towards the major class, i.e., non-drug-abuse risk behavior tweets. From the angle of classifiers, SVM model achieves the best overall performance. Random Forest has slightly less average accuracy than the SVM model, but worse recall and F1-value. Furthermore, from the view of vectorization approach, it is clear that word2vec embedding outperforms term frequency and *tf-idf* in most of the cases. Several possible combinations of settings are not shown in Fig. 2 due to poor performances.

Experiment on self-taught learning models

As shown in the previous experiment, SVM model using the custom-trained word2vec embedding achieves the best performance, we decided to apply the same model structure to compare with our deep self-taught learning approaches. In this experiment, at each epoch, 10,000 machine-labeled tweets were randomly sampled and merged into the training set. Figure 3 shows the experimental results of the five self-taught models, including self-taught CNN (*st-CNN*), self-taught LSTM (*st-LSTM*), self-taught SVM (*st-SVM*), self-taught NB (*st-NB*), and self-taught RF (*st-RF*). All configurations of classifiers and vectorization methods are tested. For the sake of clarity, we only illustrate the best performing setting for each model in Fig. 3. It is clear that our proposed deep self-taught learning approaches (i.e., st-LSTM and st-CNN) outperform traditional models, i.e., st-SVM, st-NB, and st-RF, in terms of accuracy, recall, and F1-value, in all cases. Deep learning models achieve 86.53%, 88.6%, and 86.63% in terms of accuracy, recall, and F1-value correspondingly.

Experiment on vectorization settings

The impact of two different word2vec representations on the st-CNN, i.e., the custom word2vec embedding we trained from our corpus, and pre-trained Google News word2vec embedding, is shown in Fig. 4. The Google News word2vec achieves 0.1%, 0.4%, and 0.3% improvements in terms of accuracy, recall, and F1-value (86.63%, 89%, and 86.83%, respectively) compared with the custom-trained word2vec embedding. In addition, it is clear that Google News word2vec embedding outperforms the custom-trained word2vec in most of the cases. This is because the Google News word2vec embedding was trained on a large-scale corpus, which is significantly richer in contextual information, compared with our short, noisy, and sparse Twitter datasets.

An insight analysis of drug-abuse risk behavior on Twitter

To gain insights in drug-abuse risk behaviors on Twitter, we use our best performing deep self-taught learning model to annotate over three million drug-abuse-related tweets with geo-tags and perform quantitative analysis. There are 117,326 tweets classified as positive, and 3,077,827 tweets classified as negative. The positive tweets correspond to 3.67% of the whole dataset. We performed analysis from three aspects: word and phase distributions, temporal distributions, and spatial distributions.

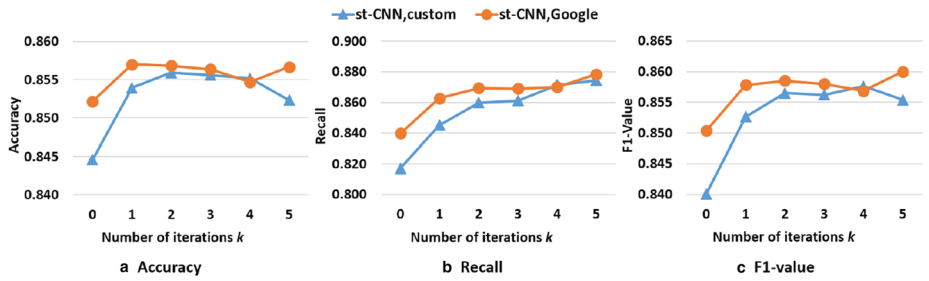

Fig. 4 Performance comparison between custom word2vec embedding and Google News word2vec embedding

a Positive tweets **b** Negative tweets

Fig. 5 Word frequency distribution

Word and phase distributions

We first visualize the top frequent words by word cloud, as shown in Fig. 5. The word distribution in positive tweets (Fig. 5a) is remarkably different from word distribution in negative tweets (Fig. 5b). In fact, drug-abuse tweets usually consist of abuse-indicating terms, and drug names, such as "blunt," "high," "smoke," "weed," "marijuana," "grass," "juic," etc. (Fig. 5a). In addition, the high concentration of dirty words, e.g., "s**t," "f**k," "as*," "bit**," etc., clearly suggests the expression patterns that the drug abusers may have (Fig. 5a). This expression pattern does not likely exist in negative tweets. Then, we further show the comparison of normalized word frequency between positive tweets and negative tweets (words from positive tweets got normalized by the number of positive tweets, and negative words by negative tweets), regarding the 25 most frequent words in positive tweets (Fig. 6) and 25 most frequent words in negative tweets (Fig. 7). Note that in Fig. 6, the *y*-axis is clipped at 0.25, which is the value of word "weed", while the word "smoke" has the normalized frequency of 0.44. These two figures further show that: (1) positive-frequent words are more likely to have lower normalized frequency in negative tweets, and vise versa and (2) some ordinary words, i.e., "go", "want", "day", and "good", still share similar normalized frequency between positive and negative tweets.

Hashtags also play an import role in the Twitter sphere as a way for users to: (1) to express their opinion more clearly and (2) to improve information sharing efficiency. Tweets that share same Hashtags can be grouped together and easily found, while popular Hashtags can make the tweets more visible to wider audience. Table 3 shows the most frequent Hashtags in positive tweets and negative tweets. It is clear that the Hashtags in positive tweets are almost exclusively related to drug abuse, while the Hashtags in negative tweets cover much wider range of topics.

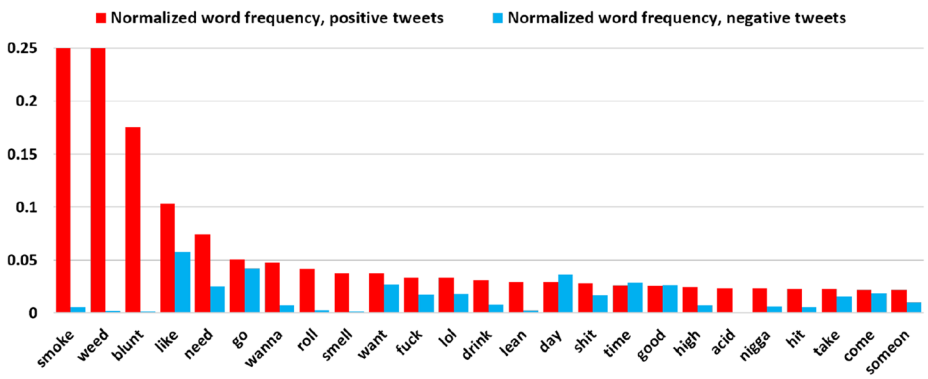

Fig. 6 Normalized frequency of top 20 frequent words in positive tweets, compared with in negative tweets

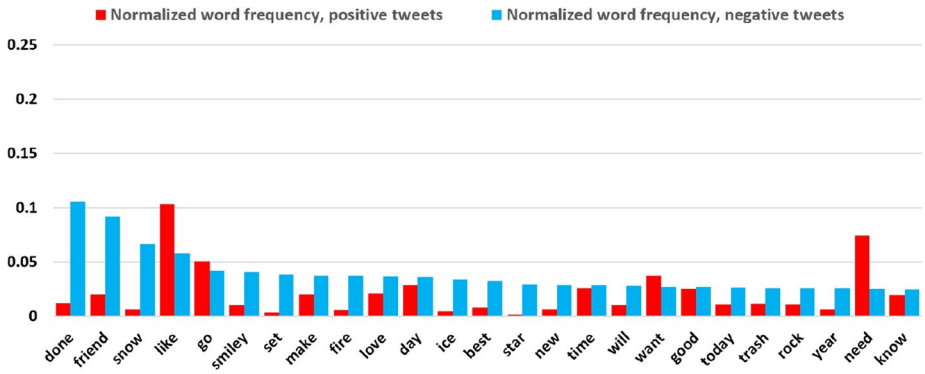

Fig. 7 Normalized frequency of top 20 frequent words in negative tweets, compared with in positive tweets

Table 3 Most frequent Hashtags in positive tweets and negative tweets

Positive tweets	#weed, #smoke, #cannabis, #marijuana, #glassofig, #scientificglass, #WeedFirm, #maryjane, #dabs, #kush, #3wordsbetterthanIloveyou, #MarijuanaFunFacts, #pot, #dank, #high, #thc, #stoner, #blunt, #highlife, #AcademyAward, #OscarNominations, #ganja, #waterpipes, #np, #herblife
Negative tweets	#job, #snow, #hiring, #photo, #traffic, #CareerArc, #NBAVote, #Simon, #winter, #jobs, #Hospitality, #peace, #WomensMarch, #love, #Toronto, #Trump, #STAR, #nowplaying, #Orlando, #AZ, #np, #Veterans, #Retail, #SoundCloud, #nyc, #Inauguration, #cat, #weather, #MAGA

Table 4 Drug name and abuse behavior co-occurrence frequency differences between positive tweets and negative tweets

Combo	Pos_count	Neg_count	Ratio_diff (%)	Relative_ratio (%)
Trash high	1131	1387	0.9189	1166.04
Acid trip	547	239	0.4585	1863.66
Acid drop	256	168	0.2127	1603.13
Glass amp	374	3472	0.2060	171.11
Acid take	222	167	0.1838	1509.61
Lean amp	280	2391	0.1610	192.55
Coke high	195	186	0.1602	1343.14
Coke take	185	512	0.1410	646.57
Lean hit	180	745	0.1292	446.33
Acid amp	162	367	0.1262	761.96
Molly pop	160	328	0.1257	823.11
Acid hit	132	138	0.1080	1278.34
Lean pop	121	118	0.0993	1327.49
Acid use	115	238	0.0903	817.21
Shrooms trip	105	55	0.0877	1751.49
Lean high	108	136	0.0876	1147.54
Lean use	159	1479	0.0875	170.62
Blow high	125	675	0.0846	337.93
Upper high	112	382	0.0830	537.16
Dope high	106	509	0.0738	383.46
Coke amp	137	1413	0.0709	146.07
Acid high	65	57	0.0535	1402.44
Coke snort	82	571	0.0513	251.20
Molly amp	88	777	0.0498	183.80
Crack hit	108	1360	0.0479	104.18

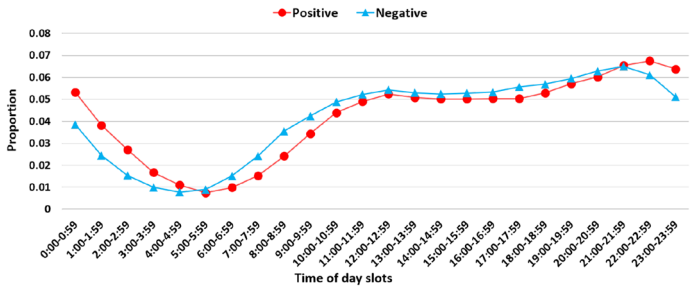

Fig. 8 Time of day distribution comparison between positive tweets and negative tweets

Table 5 Chi-square test of time of day distribution

Type	Chi square	*p*-value
All day	46.467257	0.002615305(**)
Day time (8 a.m. to 6 p.m.)	6.87318202	0.650321116
Night time (6 p.m. to 8 a.m.)	39.5940753	0.000160637(***)

(**) $p < 0.01$, (***)$p < 0.001$

Finally, for word and phase analysis, we extract the co-occurrence frequencies of combinations of drug name and drug-abuse behavior. For each combination, we count the number of positive tweets and negative tweets that contain all words in that combination, then sort it by the absolute difference of normalized frequency between positive tweets and negative tweets. Table 4 shows the top 25 observed combinations. The "*Relative_ratio*" column is showing the ratio that the combination appears in positive tweets over the appears in all tweets. This analysis spots the more frequently used drug-abuse risk behavior indication word combinations, which will support further data collection.

Temporal analysis

To examine if there are different time patterns for positive tweets and negative tweets to be posted, we extract the local posting time of each tweet, then perform 1-h interval binning. As shown in Fig. 8, where x-axis are time slots, and y-axis is the proportion (normalized count) of tweets. The results shown in Fig. 8 are very interesting. The time patterns are obviously different between positive tweets and negative tweets. In fact, the Chi-square test results on the data in Fig. 8 shown in Table 5 clarifies that the pattern differences are significant for the time frames of 'All day' and 'Night time.' This result shows a very plausible phenomenon that tweets with drug-abuse risk behaviors are more active in night time than in day time.

Spatial analysis

The geo-location information tagged in tweets is very useful for capturing the distribution of drug-abuse risk behaviors. The geo-tagging information on Twitter usually comes in two forms: GPS coordinates, or a "Place Object" associated with the tweet. We first visualize geo-distribution of the positive tweets by plotting each geo-tag across the

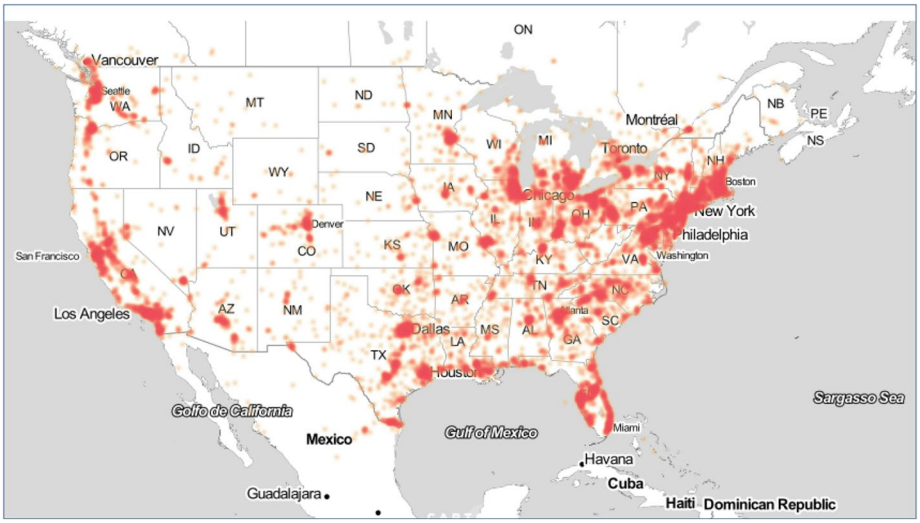

Fig. 9 Dot map of positive tweets across the United States

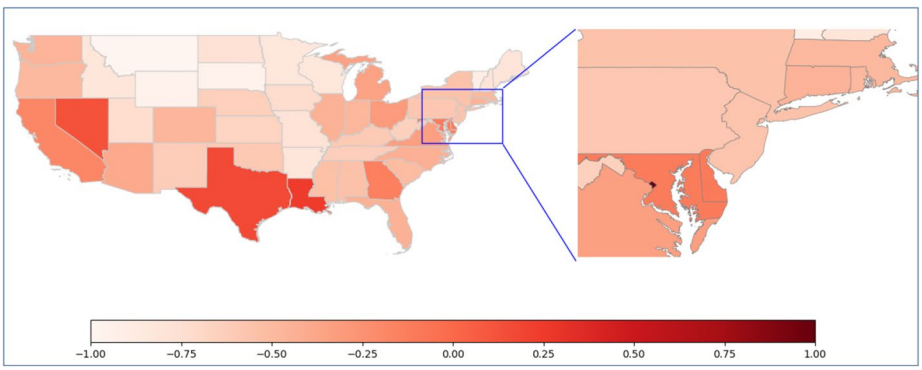

Fig. 10 Number of positive tweets, per state, normalized by state population of 12 or older

continental United States in Fig. 9. By making this fine granularity, we can confirm that the collected tweets generally follow the population distribution. Then, we aggregate the geo-tags into state level, normalized with state's population of age group 12 or older, and draw Fig. 10 with the numbers scaled to $[-1,1]$. From Fig. 10, we can see that the District of Columbia has an extremely high ratio of positive tweets, follow by Louisiana, Texas, and Nevada that have relative high rate. Other states with high rate including California, Georgia, Maryland, and Delaware. Furthermore, the distribution of other states' data showing that the collected tweets align relatively well with state-level population distribution.

The other spatial analysis we perform is the alignment between our state-level counts of positive tweets, normalized with state population, and the 2016–2017 National Survey on Drug Use and Health (NSDUH) survey data. Here, the normalization is meant to decorrelate the count of tweets from the population of each state, and is done by simply dividing the count of positive tweets by the population (2017 census estimation) for each state. We choose to perform normalization with population for two reasons: (1) we

have little to no control of the sampling process, in terms of geo-location distribution, when crawling data from Twitter, which means the bias is unavoidable and uncontrollable and (2) thus, the state population figures are more reliable, stable, and representative. NSDUH is a creditable source of drug-abuse-related population scale estimation. If our Twitter data can align with the reliable survey data, we can argue that the Twitter-based studies have the prediction power that should not to be ignored. By computing the Pearson's R between the normalized number of tweets and the NSDUH prevalence rate, over the same age group (12 or older), it is surprising to find that in our study even without further categorization, the Twitter data are significantly correlated ($p < 0.05$) with some of the most important categories in the NSDUH study: (1) "Illicit Drug Use Other Than Marijuana in the Past Month" ($r = 0.387$); (2) "Cocaine Use in the Past Year" ($r = 0.421$); (3) "Methamphetamine Use in the Past Year" ($r = -0.372$); (4) "Pain Reliever Use Disorder in the Past Year" ($r = -0.375$); and (5) "Needing But Not Receiving Treatment at a Specialty Facility for Illicit Drug Use in the Past Year" ($r = 0.336$). We argue that when large quantity of Twitter data is available, we can perform more detailed and creditable studies on the population scale.

Discussion and limitations

According to our experimental results, our deep self-taught learning models achieved promising performance in drug-abuse risk behavior detection in Twitter. However, many assumptions call for further experiments. First, how to optimize the classification performance by exploring the correlations among parameters and experimental configurations. For instance, for SVM and RF models, unigram feature works better than n-gram feature on term frequency; however, for *tf-idf*, it is the opposite situation. Second, the pre-trained Google News word2vec embedding performs better than the custom-trained word2vec embedding may also be situational. These findings indicate the necessity of leveraging size and quality of the training data for training word embedding, given that the available data may better fit the classification task but be short in quantity. Nevertheless, among the measures, recall receives a more significant boost than accuracy and F1-value. We may argue that the proposed self-taught algorithm helped correcting the bias in the classifiers caused by the imbalanced nature of the training dataset. However, more experiments need to be conducted to verify this interesting point.

Future research

The study we presented in this paper can be improved in many ways. Here, we elaborate several of the future research directions. First, we plan to incorporate the well-trained classifier into a real-time drug-abuse risk behavior monitoring and analysis system that aims at providing community-level stakeholders with timely accessible detection results for supporting their efforts, such as recovery services and public educations, on combating the opioid crisis. Second, we can utilize more information that can be extracted from tweets, such as user tweeting history, user demographic attributes, and user interactions, to further improve the model in terms of performance, scope, and credibility. Third, the extra information that we extract further

enables the analysis of connections among users and tweets, on both social network plane and geospatial network plane, which can help to acquire knowledge regarding how the drug trend propagates through both planes. Last but not least, we may expand the study to other major online social media platforms, i.e., Reddit and Instagram, and more specialized online forum Blulelight.

Conclusion

In this paper, we proposed a large-scale drug-abuse risk behavior tweet collection mechanism based on supervised machine-learning and data crowd-sourcing techniques. Challenges came from the noisy and sparse characteristics of Twitter data, as well as the limited availability of annotated data. To address this problem, we propose deep self-taught learning algorithms to improve drug-abuse risk behavior tweet detection models by leveraging a large number of unlabeled tweets. An extensive experiment and data analysis were carried out on three million drug-abuse-related tweets with geo-location information, to validate the effectiveness and reliability of our system. Experimental results shown that our models significantly outperform traditional models. In fact, our models correspondingly achieve 86.53%, 88.6%, and 86.63% in terms of accuracy, recall, and F1-value. This is a very promising result, which significantly improves upon the state-of-the-art results.

Further data analysis gain insights into the expression patterns and the geo-distribution that the drug abusers may have on Twitter. For example, the words and phrases used in drug-abuse risk behavior-positive tweets have distinctive frequencies that can be used in data collection to improve the quality of raw data. The uneven geographical distribution of tweets makes it appealing to perform further analysis that associates tweets with other geographical data.

Abbreviations

NSDUH: National Survey on Drug Use and Health; CDC: Centers for Disease Control and Prevention; CNN: convolution neural networks; LSTM: long–short-term memory; DAWN: drug-abuse warning network; SVM: support vector machine; NB: Naive Bayes; RF: Random Forest; AMT: Amazon Mechanical Turk; URL: uniform resource locator; HTML: hypertext markup language; *tf*: term frequency; *tf-idf*: term frequency-inverse document frequency; API: application programming interface; GPU: graphics processing unit; GPS: global positioning system; NDTA: National Drug Threat Assessment.

Acknowledgements

The authors gratefully acknowledge the support from the National Science Foundation (NSF) Grants CNS-1650587, CNS-1747798, CNS-1624503, CNS-1850094, and National Research Foundation of Korea NRF-2017S1A3A2066084.

Authors' contributions

DD and RJ have a significant and intellectual advice to shape the extension with spatial data analysis, combining with offline data collected from NDTS [46]. Thanks to XY, KD, and DK with their expertises in geospatial data analysis and drug-abuse behaviral analysis, all the results in the analysis were verified and irrelevant results were eliminated. KD and DK are experts in statistics, especially in substance abuse, verified all the results in Fig. 8 and Table 5. KD and DK also verified the statistical results in the comparison between our data with the NDTS data [46] to eliminate uncertain results. Upon that, we discovered new interesting and statistically significant data correlations. HH and NP conducted and verified the experiment. HV contributed the data and the visualization part. SAC and JG contributed to the system development and data annotation processes. All authors read and approved the final manuscript.

Funding

The authors gratefully acknowledge the support from the National Science Foundation (NSF) Grants CNS-1650587, CNS-1747798, CNS-1624503, CNS-1850094, and National Research Foundation of Korea NRF-2017S1A3A2066084.

Availability of data and materials

The data are available upon request, following the data privacy policy of Twitter.

Competing interests

The authors declare that they have no competing interests.

Author details
[1] New Jersey Institute of Technology, University Heights, Newark 07102, USA. [2] City University of New York, 2800 Victory Blvd, Staten Island 10314, USA. [3] The City College of New York, 160 Convent Ave, New York 10031, USA. [4] Kent State University, 800 E. Summit St., Kent 44242, USA. [5] University of Oregon, 1585 E 13th Ave., Eugene 97403, USA.

Received: 2 June 2019 Accepted: 17 October 2019
Published online: 06 November 2019

References

1. U.S. Department of Health and Human Services: HHS acting secretary declares public health emergency to address national opioid crisis. 2017.
2. Substance Abuse and Mental Health Services Administration, U.S. Department of Health and Human Services: key substance use and mental health indicators in the United States: results from the 2016 National Survey on Drug Use and Health. 2018. http://datafiles.samhsa.gov. Accessed 20 May 2019.
3. Substance Abuse and Mental Health Services Administration, U.S. Department of Health and Human Services: key substance use and mental health indicators in the United States: results from the 2015 National Survey on Drug Use and Health. 2018. http://datafiles.samhsa.gov. Accessed 20 May 2019.
4. National Institute on Drug Abuse, U.S. National Institutes of Health: overdose death rates. 2018.
5. The Gun Violence Archive: 2015 Gun Violence Archive. 2018. http://www.gunviolencearchive.org/past-tolls. Accessed 20 May 2019.
6. Aslam S. Twitter by the numbers. 2018. http://www.omnicoreagency.com/twitter-statistics/. Accessed 20 May 2019.
7. Signorini A, Segre AM, Polgreen PM. The use of twitter to track levels of disease activity and public concern in the us during the influenza a H1N1 pandemic. PLoS ONE. 2011;6(5):19467.
8. Aphinyanaphongs Y, Lulejian A, Brown DP, Bonneau R, Krebs P. Text classification for automatic detection of e-cigarette use and use for smoking cessation from twitter: a feasibility pilot. In: Biocomputing 2016: proceedings of the Pacific symposium. 2016. p. 480–91.
9. Bosley JC, Zhao NW, Hill S, Shofer FS, Asch DA, Becker LB, Merchant RM. Decoding twitter: surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation. 2013;84(2):206–12.
10. Chary M, Genes N, McKenzie A, Manini AF. Leveraging social networks for toxicovigilance. J Med Toxicol. 2013;9(2):184–91.
11. Hossain N, Hu T, Feizi R, White AM, Luo J, Kautz H. Precise localization of homes and activities: detecting drinking-while-tweeting patterns in communities. In: Tenth international AAAI conference on web and social media. 2016.
12. Myslín M, Zhu S-H, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Medical Internet Res. 2013;15(8):e174.
13. Sarker A, et al. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. Drug Saf. 2016;39(3):231–40.
14. 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.
15. 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.
16. 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.
17. 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.
18. Hu H, Phan N, Geller J, Vo H, Manasi B, Huang X, Di Lorio S, Dinh T, Chun SA. Deep self-taught learning for detecting drug abuse risk behavior in tweets. In: International conference on computational social networks. 2018. p. 330–42.
19. LeCun Y, Bottou L, Bengio Y, Haffner P, et al. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
20. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
21. Johnston L, National Institute on Drug Abuse. Monitoring the future: National survey results on drug use, 1975–2004, vol. 1. 2005.
22. Brookoff D, Campbell EA, Shaw LM. The underreporting of cocaine-related trauma: drug abuse warning network reports vs hospital toxicology tests. Am J Public Health. 1993;83(3):369–71.
23. Kessler DA, Natanblut S, Kennedy D, Lazar E, Rheinstein P, Anello C, Barash D, Bernstein I, Bolger R, Cook K, et al. Introducing medwatch: a new approach to reporting medication and device adverse effects and product problems. JAMA. 1993;269(21):2765–8.
24. Meng H-W, Kath S, Li D, Nguyen QC. National substance use patterns on twitter. PLoS ONE. 2017;12(11):1–15. https://doi.org/10.1371/journal.pone.0187691.
25. Ding T, Bickel WK, Pan S. Social media-based substance use prediction. arXiv preprint arXiv:1705.05633. 2017.
26. Simpson SS, Adams N, Brugman CM, Conners TJ. Detecting novel and emerging drug terms using natural language processing: a social media corpus study. JMIR Public Health Surveill. 2018;4(1):2.
27. Phan NH, Chun SA, Bhole M, Geller J. Enabling real-time drug abuse detection in tweets. In: 2017 IEEE Int. Conf. Data Eng. (ICDE). 2017. p. 1510–4.

28. Coloma PM, Becker B, Sturkenboom MC, van Mulligen EM, Kors JA. Evaluating social media networks in medicines safety surveillance: two case studies. Drug Saf. 2015;38(10):921–30.
29. Hu H, Moturu P, Dharan K, Geller J, Iorio S, Phan H, Vo H, Chun S. Deep learning model for classifying drug abuse risk behavior in tweets. In: 2018 IEEE international conference on healthcare informatics (ICHI). IEEE; 2018. p. 386–7.
30. Kong C, Liu J, Li H, Liu Y, Zhu H, Liu T. Drug abuse detection via broad learning. In: International conference on web information systems and applications. Berlin: Springer; 2019. p. 499–505.
31. Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inform Assoc. 2019;. https://doi.org/10.1093/jamia/ocz156.
32. Mahata D, Friedrichs J, Shah RR, Jiang J. Detecting personal intake of medicine from twitter. IEEE Intell Syst. 2018;33(4):87–95.
33. Zhang Y, Fan Y, Ye Y, Li X, Winstanley EL. Utilizing social media to combat opioid addiction epidemic: automatic detection of opioid users from twitter. In: Workshops at the thirty-second AAAI conference on artificial intelligence. 2018.
34. Li J, Xu Q, Shah N, Mackey TK. A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study. J Med Internet Res. 2019;21(6):13803.
35. Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th international conference on machine learning. 2007. p. 759–66.
36. Bengio Y, et al. Learning deep architectures for AI, foundations and trends. Mach Learn. 2009;2(1):1–127.
37. Weston J, Ratle F, Collobert R. Deep learning via semi-supervised embedding. In: Proceedings of the 25th international conference on machine learning. 2008. p. 1168–75.
38. Bettge A, Roscher R, Wenzel S. Deep self-taught learning for remote sensing image classification. 2017. arXiv preprint arXiv:1710.07096.
39. Dong X, Meng D, Ma F, Yang Y. A dual-network progressive approach to weakly supervised object detection. In: Proceedings of the 25th ACM international conference on multimedia. 2017. p. 279–87.
40. Gan J, Li L, Zhai Y, Liu Y. Deep self-taught learning for facial beauty prediction. Neurocomputing. 2014;144:295–303.
41. Yuan Y, Liang X, Wang X, Yeung D-Y, Gupta A. Temporal dynamic graph lstm for action-driven video object detection. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 1801–10.
42. U.S. National Institute on drug abuse: commonly abused drugs. 2018.
43. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. 26th NIPS, vol. 2. 2013. p. 3111–9.
44. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. 2013. p. 245–51.
45. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23.
46. U.S. Department of Drug Enforcement Administration: National Drug Threat Assessment. 2018.

Publisher's Note