Original Paper

What are Users Across Different Age Groups Talking About Drugs on Reddit?

Qiudan Li¹, PhD; †Can Wang¹,², BE; †Hejing Liu¹, BE; †Daniel Dajun Zeng¹,²,³, PhD

¹The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
²University of Chinese Academy of Sciences, Beijing, China
³Department of Management Information Systems, Eller College of Management, The University of Arizona, Tucson, AZ, United States

*Corresponding Author:
Daniel Dajun Zeng, PhD
Department of Management Information Systems
Eller College of Management
The University of Arizona
McClelland Hall 430
1130 E Helen St
Tucson, AZ, 85721
United States
Phone: 1 520 621 4614
Fax: 1 520 621 2433
Email: zeng@email.arizona.edu
†Equal contributors
Abstract

**Background:** Drug abuse is a severe challenge for the United States. Social media such as Reddit has become one of the most efficient platforms for drug users to share experiences and communicate with each other. By analyzing users' perceptions and behavior patterns concerning causes and symptoms of drug usage on social media, public health researchers and agencies can have innovative real time situational awareness and surveillance capabilities to tackle the drug abuse crisis.

**Objectives:** This paper aims to develop a social media-based approach to analyze users' causes and symptoms of drug usage across different age groups, then deeply gain insights into users' behavior pattern.

**Methods:** We collected 163,610 posts on the Reddit /r/drugs community from February 2008 to December 2017. Firstly, topics for drugs, causes, and symptoms based on word vector learning (word2vec) were extracted. Then, we designed a method to automatically extract age information from posts. Finally, the relationships between age and drugs, causes, and symptoms were established.

**Results:** We found that: (1) Drug topics contained 6 categories including alcoholic, tobaccos, prescriptions, hallucinogens, ecstasies, and other highly
**addictive drugs.** Tobaccos (n=11,947), hallucinogens (n=9244), and other highly addictive drugs (n=5733) were the most discussed drug categories. (2) 9646 (65% of the total posts that had age information) posts on Reddit were sent by users between 15 and 25 years old. There was relevance between age and drug addiction. People tended to change their drug usage from a primary drug (e.g. marijuana) to other highly addictive drugs (e.g. heroin) as they aged. The age groups age<15 and 15<=age<20, 20<=age<25 had very similar drug topic patterns, where the Pearson correlation coefficient (PCC) between age<15 and 15<=age<20 is .994 (P<.001), the Pearson correlation coefficient (PCC) between age<15 and 20<=age<25 is .985 (P<.001). Similar phenomena were also obtained for age groups 20<=age<25 and 25<=age<30, where the Pearson correlation coefficient (PCC) is .989 (P <.001). (3) The reasons users take different types of drugs can be classified as curious, anxious and pain. The most frequently mentioned cause is curiosity about a novel experience (1215, 1215/1848), which may due to the characteristics of young users. For the curiosity cause, the ratios of posts discussing tobaccos and hallucinogens were 0.5045 (613/1215) and 0.3975 (483/1215) respectively. (4) The symptoms caused by taking drugs can be classified into 6 topics, including hallucination, comfort, lose_control, affect_brain, anxious and pain. Anxiety was the most popular symptom mentioned among the posts (574, 574/1096). The number of posts reporting
anxious symptom caused by drugs is a little larger than that of posts discussing curing anxiety by drugs (n=558).

**Conclusions:** This is the first study aimed at deeply gaining insights into users’ behavior patterns, which can help public health researchers and agencies provide personalized health regulatory services for people at different ages.

**Key words**

Perceptions, Behaviors, Drug Topic, Reddit, Social Media
Introduction

Drug abuse is an increasingly significant public health problem. According to a behavioral health overview by SAMHSA [1] in 2015, about 12.5 million users aged 12 and over reported misusing prescription pain relievers, and about 828,000 reported using heroin. Social media platforms such as Twitter, Facebook, Reddit, and MedHelp, etc. are used by young adults to communicate with each other and share their drug use experience [8]. Based on Infodemiology and Infoveillance proposed by Eysenbach [2], these communication and publication patterns on social media have important implications for a population's health and public policy. By analyzing users’ perceptions and behavior patterns concerning causes and symptoms of drug use on social media, public health researchers and agencies can have innovative, real-time situational awareness and surveillance capabilities to tackle the drug abuse crisis.

Due to its characteristics of anonymity, Twitter provides a platform for adolescents to exchange information about prescription drug abuse [3]. [4] conducted a survey to understand young adults’ exposure to pro-alcohol and marijuana-related content on Twitter. [5] assessed the content of tweets and demographics of consumers who follow a popular Twitter handle and underscored the need for surveillance efforts to monitor the pro-marijuana content reaching young people on Twitter. [6] examined marijuana-related content in Twitter, especially that tweeted by adolescent users, to examine any differences in marijuana-related message content.
before and after two states legalized recreational marijuana. [3] demonstrated the use of Twitter posts as a way to examine Adderall abuse among a sample of college students in the United States. Recently, deep-learning-based methods have been proved to be effective in learning relationships and word vector representations. [9] trained a continuous bag-of-words (CBOW) model of distributed word vector embeddings to identify and track emerging drug term trends on Twitter, particularly with reference to terms for marijuana. [10] proposed a convolutional recurrent neural network to perform adverse drug reaction classification on Twitter and MEDLINE case reports. [11] showed that big health data analytics can be fruitfully leveraged with deep-learning techniques to uncover new and actionable insights about medication safety.

Compared with Twitter, web forums also contain many conversations about nonmedical drug use. [7] showed that content analysis of social listening data could be utilized to identify posts discussing potential misuse or nonmedical use of bupropion and two comparators, amitriptyline and venlafaxine. Reddit is one of the most popular forums in the world. As of June 28, 2015, Reddit had 163,966,958 unique visitors hailing from over 212 different countries, viewing a total of 7,086,828,967 pages according to their published statistical data [12]. Recently, Reddit is being used as a new data source for public health research. [13] used Reddit to study mental health, and [14] used Reddit to track the outbreak of Ebola in 2014. [27] suggested that Reddit could be used for data mining and analysis of e-cigarette-related content. [15] utilized Reddit to study e-juice components. [16]
identified topics for e-cigarette user-generated contents from multiple social media platforms. On Reddit, users are allowed to create a community (called a “subreddit”) where they can discuss interesting topics with each other. For instance, there are many publicly available posts about user experiences, feelings, and symptoms caused by different drugs, which have the potential to influence drug-related attitudes, choices, and behaviors. [25] analyzed the severity of posters’ opioid use disorder based on a few posts on Reddit. However, there are no published studies to date that have systematically mined users' drug experiences and symptoms utilizing data from Reddit. This study aims to gain a systematic understanding of causes and symptoms of prescription drug usage by analyzing posts from the drugs subreddit on Reddit. The data-driven findings mined from social media could benefit regulatory agencies like the FDA, enabling them to obtain a better understanding of drug use and formulate personalized policy.

Methods

Figure 1 shows the framework for analyzing users’ posts on the Reddit drugs community. It consists of three components: Data Collection and Preprocessing, Data Analysis, and Results.
Data Collection

Using the Reddit Application Programming Interface (API), we collected 163,610 posts including user name, published time, content, and replies from the subreddit /r/drugs published by 79,786 users from February 2008 to December 2017.

To better understand the perceptions and behaviors of users of different age groups, age information extraction is necessary. Firstly, regular expressions were designed to automatically extract age information from posts; for example, age 18 is extracted from posts containing patterns like "18 years old" or "I'm 18." We divided

---

1 https://www.reddit.com/r/Drugs/
posts into 5 age groups: (1) age < 15 (juveniles), (2) 15 <= age < 20 (teenagers), (3) 20 <= age < 25 (young adults and college-aged people), (4) 25 <= age < 35 (adults), and (5) age >= 35.

Data Analysis

To understand behaviors concerning what and why users in different age groups take drugs and their symptoms after drug usage, it is necessary to firstly capture the topics about drugs, causes, and symptoms, and then establish the relationships between age information and the identified topics.

However, the large amount of posts containing high-level noise, such as heavy use of slang, misspellings, etc., makes it extremely difficult for users to annotate and extract topics manually. Therefore, the key challenge of identifying topics is to learn latent semantic information among words and solve the semantic fuzziness. Deep-learning methods are representation-learning methods, which can automatically discover multiple levels of representations from raw data. They are considered promising methods for various tasks including topic classification and knowledge extraction [23]. Word2vec is a deep-learning-inspired method that attempts to understand meanings and semantic relationships among words. It learns vector representations of words using continuous bag-of-words (CBOW) and SkipGram [24]. By learning the vector representation of words, the words with similar semantics are mapped into adjacent locations in the vector space. We used a 2-step process to identify topics and their words. 3 annotators were invited to code the top 600 words with high frequency until an acceptable interrater reliability score was reached, kappa coefficient was above 0.7, which indicated substantial or good agreement, initial word dictionaries for drugs, causes and symptoms were obtained; then word2vec was
adopted to learn effective representation of the low-dimensional word vectors to enrich the word dictionary.

To build a mapping from age groups to identified topics, we calculated the percentage of posts that contain the age group information and topic category's words. If one post contained more than one category, we chose the category the post mentioned most. The pearson coefficient was calculated to further analyze the topic differences among different age groups.

Results

1) Drug Topics and Drug Usage Analysis for Different Age Groups

Drug Topics

According to the strength of addiction, drug topics are classified into six categories: alcoholic, tobaccos, prescriptions, hallucinogens, ecstasies, and other highly addictive drugs. Table 1 shows the topic categories and associated keywords, with expanded keywords shown in italics. We selected commonly used keywords and abbreviations as initial keywords, and then used word2vec to obtain more comprehensive topic keywords including variations of drug names and misspellings. The first topic was about alcohol-related drugs, and word2vec identified related abbreviations and misspellings such as alc, alchol, etc. The second topic was discussing tobacco-related drugs. Marijuana was a common tobacco drug, and by using word expansion, we can identify words such as slang, pot, cannabis, and the
abbreviation *mj*. Words identifying how users used the drug were also identified, such as *cigarette* and *cig*. Topic 3 was about prescription drugs, which are used because of psychological or physiological diseases. Examples include *kratom*, a drug for pain relief; drugs helping focus attention such as *potentiator, adderall, Vyvanse, Dexedrine, and concerta*; and drugs for depression treatments including *Tianeptine* and *gabapentin*. Drugs for depression were also used with other drugs such as *addy* (containing adderall and vyvanse) due to a variety of physical or mental illnesses.

Topic 4 discussed hallucinogens, which make users produce illusions. *Mushroom, LSD* (also called acid), and *DMT* (dimethyltryptamine) were commonly used drugs. Other LSD nicknames such as *Lucy* and *cid*, and various kinds of mushrooms, such as *Cyanescens, cubensis, and psilocybe*, were obtained by word2vec. Topic 5 was about ecstasy drugs, including MDMA aliases, such as *molly* and *ecstasy*, similar drugs to *MDMA, mda*, and abbreviations of MDMA and ecstasy such as *md*. Topic 6 contained a variety of highly addictive drugs such as *meth* and *heroin*. The associated potent drug *cocaine* and its nicknames such as *coke* and *crack* were also found by word2vec.

<table>
<thead>
<tr>
<th>Drug topic categories</th>
<th>Key words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Alcoholic</strong></td>
<td>Alcohol</td>
</tr>
<tr>
<td></td>
<td><em>alcohol, booze, alc, alchol</em></td>
</tr>
<tr>
<td>2. <strong>Tobaccos</strong></td>
<td>Weed, marijuana</td>
</tr>
<tr>
<td></td>
<td><em>Pot, bud, cannabis, mj, cigs, cigarettes</em></td>
</tr>
<tr>
<td>3. <strong>Prescriptions</strong></td>
<td>Adderall,DXM</td>
</tr>
<tr>
<td></td>
<td><em>Kratom, Tianeptine, gabapentin, potentiator, amatine, maeng, Vyvanse, Ritalin, Dexedrine,</em></td>
</tr>
</tbody>
</table>
4. Hallucinogens
Shrooms, acid, LSD, DMT
Lucy, cid, shroom, mushroom, Cyanescens, cubensis, psilocybe

5. Ecstasies
Mdma
Molly, ecstasy, mda, md, xtc

6. Other highly addictive drugs
Meth, heroin, cocaine
crack, methamphetamine, heroine, coke

Drug usage analysis

By analyzing users' demographic characteristic of age for different categories of drugs, we can gain deeper insights into users' preferences and then help regulators design more personalized strategies.

Age information was extracted from 15,545 posts. The number of posts in the groups <15, [15,20), [20,25), [25,35), and >=35 are 3431, 7326, 2330, 1507, and 951 respectively. To analyze the drug use patterns of users across different age groups, the number of posts that contain keywords of each drug category and age group information were counted. Tobaccos (n=11,947), hallucinogens (n=9244), and other highly addictive drugs (n=5733) were the most discussed drug categories, while alcoholic (n=3295) and ecstasies (n=1204) were the least mentioned. For each drug category mentioned above, the most active users are between 15 and 20 (ratio of posts: 6253/11,947, 4754/9244, 2280/5733, 1524/3295, 560/1204). The detailed numbers can be seen in Supplemental File 1. Figure 2 presents the ratios of posts for each drug category across different age groups. For users below 35 years old, tobaccos, hallucinogens, and other highly addictive drugs were the most discussed drug categories, while for users above 35 years old, other highly addictive drugs,
hallucinogens, and tobaccos were the most popular drugs. The number of posts for other highly addictive drugs showed a rising trend as people aged, while the trend of the other five drug categories first increased and then decreased. It can be seen that tobacco products, including cannabis and weed, were less addictive and more attractive to users who were younger, less addicted, or new users, while highly addictive drugs were more attractive to users who were older, drug addicts, or took drugs for a long time.

Figure 2 Relationship between age and drug topics

In order to further analyze the topic differences among different age groups, the pearson coefficient was calculated. The age groups age<15 and 15<=age<20, 20<=age<25 had very similar drug topic patterns, where the Pearson correlation coefficient (PCC) between age<15 and 15<=age<20 is .994, the Pearson correlation coefficient (PCC) between age<15 and 20<=age<25 Pearson correlation
coefficient (PCC) is .985. Similar phenomena were also obtained for age groups
20<=age<25 and 25<=age<30, where the Pearson correlation coefficient (PCC) is .989. Note that all P values for the abovementioned PCC values are less than .001.

2) Topics and Trends Analysis of Causes and Symptoms

Topics of causes

The reasons users take different types of drugs can be classified as curious, anxious, and pain. Supplemental File 2 shows the causes topic categories and their keywords, with the initial words shown on the first line and their extended words on the second line in italics. The curious reason can be described as: when users wanted to try something new or became addicted to drugs, they asked for help and advice about the use or side effects of certain types of drugs. Users expressed their curious and excited feelings using keywords such as curiosity, excitement, and awesome. Word2vec also identified some slang and misspellings about pursuit of novel experience including fuuuun and solotripping. The anxious reason was about seeking medical help for depression and anxiety. Users may feel paranoia or even have panic attacks. Keywords such as demotivation, tiredness, and paranoid were used to describe and explain the paranoia feelings. Some misspellings such as tiredness and axious were detected. The pain reason was about using drugs to relieve pain caused by diseases and accidents, such as earache, crampy, backache, soreness, etc.

Trends Analysis of Causes
Figure 3 presents the number of posts that mentioned each reason. The results show that *curious* (n=1215) is the most popular reason mentioned in posts, followed by *anxious* (n=558) and *pain* (n=75).

**Figure 3 Causes for users of different age**

<table>
<thead>
<tr>
<th>Age</th>
<th>Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>age&lt;15</td>
<td>277</td>
</tr>
<tr>
<td>15≤age&lt;20</td>
<td>641</td>
</tr>
<tr>
<td>20≤age&lt;25</td>
<td>292</td>
</tr>
<tr>
<td>25≤age&lt;30</td>
<td>63</td>
</tr>
<tr>
<td>30≤age&lt;40</td>
<td>112</td>
</tr>
<tr>
<td>40≤age&lt;50</td>
<td>89</td>
</tr>
<tr>
<td>50≤age&lt;60</td>
<td>12</td>
</tr>
<tr>
<td>60≤age&lt;70</td>
<td>14</td>
</tr>
<tr>
<td>70≤age&lt;80</td>
<td>6</td>
</tr>
<tr>
<td>80≤age&lt;90</td>
<td>12</td>
</tr>
<tr>
<td>90≤age&lt;100</td>
<td>5</td>
</tr>
</tbody>
</table>

To explore the relationships between causes and drugs, we counted the number of posts containing keywords about causes and drugs. It can be seen from Figure 4 that both *tobaccos* and *hallucinogens* were main causes and they had bigger impacts on *curious*. For all the causes, *tobaccos* accounted for a relatively higher proportion. For example, the ratios of *curious* posts discussing *tobaccos* and *hallucinogens* were 0.5045 (613/1215) and 0.3975 (483/1215).

**Figure 4 Relationship between causes and drugs**
Topics of Symptoms

The symptoms caused by taking drugs can be classified into 6 topics:

*hallucination, comfort, lose_control, affect_brain, anxious, and pain.* Hallucination referred to users’ perceptual experience in the absence of a corresponding objective stimulus, that is to say hallucinations. Comfort topic was about feelings of comfort, such as relaxed, anxiety relief, turning talkative, and sexually excited. These 2 topics were both short-term experiences and people may feel very high during this period. Lose_control can be described as when drug users lose control of their bodies such as having seizures and epilepsies. This symptom may be caused by overdose or withdrawal. Affect_brain was about side effects or brain damage. Users may feel sleepy or less conscious after a period of taking drugs. Anxious and pain had the same keywords as those used in the causes categories. That is because these two topics can be both reasons for drug use and symptoms after drug use. When described as symptoms, anxious was about bringing more anxiety and making people feel depressed. This happened more commonly with weed and marijuana.
Pain may be a consequence of drug overdose, making users have cramps and certain soreness. These topics and their keywords are listed in Supplemental File 3.

**Trends Analysis of Symptoms**

Figure 5 presents the number of posts that mentioned each symptom. The results showed that *anxious* \((n=574)\) was the most popular symptom mentioned in posts, followed by *affect\_brain* \((n=158)\), *hallucination* \((n=138)\), *lose\_control* \((n=108)\), *comfort* \((n=68)\), and *pain* \((n=50)\). The number of posts reporting anxious symptoms caused by drugs is slightly larger than that of posts discussing curing anxiety by drugs \((n=558)\). The number of posts reporting pain symptoms caused by drugs is slightly smaller than that of posts discussing pain causes of drugs \((n=75)\).

Figure 5 Symptoms for users of different age

To explore the relationship between symptoms and drugs, we counted the number of posts containing keywords about symptoms and drugs. It can be seen
from Figure 6 that both tobaccos and hallucinogens can lead to all the symptoms, accounting for the largest proportion in anxious symptoms (n=314, 166).

Figure 2 Relationship between symptoms and drugs

For anxious symptoms, a user expressed that “weed makes me so anxious.” For affect_brain symptoms, some users worried that marijuana will give them brain damage. For example, a user said: “I heard from some studies that marijuana has some long-term effects on the developing brain and I’m starting to get worried if I have done some irreversible damage to my brain.” Another user shared his/her experience as: “I smoked weed when I was 12 and was convinced my brain was mush and my memory was gone.”

As for hallucinations, users discussed visual and auditory hallucinations that may be related to smoking drugs. One user said: “After smoking, I had strong hallucinations that rapidly shifted and changed colors.”
For lose control symptoms, some users talked about the reasons and results of seizures. Some users described the experiences as: "doing 500mg of mdma caused a seizure on the floor"; “the seizure put me in the hospital”; “the first times i did LSD, i was also very worried about it, because of my epilepsy”. The examples showed that inappropriate use of hallucinogens and ecstasies would probably lead to a seizure.

Discussion

Principal Findings

Reddit is among the most popular forums in the world, many users share their drug use experience with each other in the drug community, yet it hasn't been fully explored in drug use analysis research compared with other platforms (e.g. Twitter and Facebook). This study provides a systematic analysis of users’ perceptions and behaviors concerning drug use across different age groups on Reddit. Our findings suggested that in-depth research regarding causes and symptoms of young adults’ drug use may increase knowledge about why and how youth and young adults use drugs and provide a foundation for developing effective strategies.

Compared with Twitter, a corpus from a forum like Reddit has serious problems such as slang, misspellings, and invalid information. Word2vec provided an efficient way to solve these problems. This is the first use of word2vec for topic discovery on Reddit, which extended existing research that mainly used word2vec on public literature and Twitter [17,9].
We found relevance between age and drug addiction. Younger users tend to use low-addition drugs (e.g. marijuana). As the period of drug usage increases, the addiction becomes stronger, and users may increase demand for harder drugs (e.g. hallucinogens, heroin, etc.). Based on this finding, it is necessary to provide effective intervention and guidance for youth group to prevent drug abuse and drug dependency as early as possible.

Our study showed that most users on the drugs subreddit were about 20 years old, which is consistent with previous findings [3,5,18]. For example, [3] found that Twitter is most popular among young adults. Tobacco topics such as weed were discussed most frequently. Young people become intrigued with it mainly because of curiosity of a novel experience, which fits the characteristics of young adults. Previous studies on other platforms also showed that young people thought marijuana would lead to pleasant and comfortable feelings and make them relax. Other reasons for marijuana are to relieve physical and psychological agony, such as anxiety and pain issues. This could be explained by the fact that medical use of marijuana to treat severe pain and other disorders is legal in some states [19]. Our findings further explained the deep reason why marijuana has a high degree of support among young people in existing studies [20]. It is important to educate the youth group about the dangers of drug use and regulate policy to monitor drug shopping, thus decrease drug abuse trend.
Our study showed that drug use symptoms can be classified into 6 topics: *hallucination*, *comfort*, *lose_control*, *affect_brain*, *anxious*, and *pain*. According to [21], hallucinogenic drugs may reduce social motivation, bring anxiety, and alter perception of the environment, which strongly affected social behaviors and caused social withdrawal [22]. [26] listed adverse health effects of marijuana use including cognitive impairment, motor coordination impairment, diminished life satisfaction, and increased risk of chronic psychosis disorders. [23] identified pain from negative comments mentioning muscle, bone, or joint pain, and they found that people on social media focused on fewer and less-serious adverse events affecting their quality of life. Our findings suggested that the relationship between drug ingredients and symptoms should be further studied by medical and public health communities.

Figure 7 shows an application scenario of the proposed framework, which can help policy makers automatically monitor behaviors of drug use. Take symptom topic as an example, based on the extracted knowledge like: “marijuana, brain damage, negative”, “MDMA, seizure, negative”, “LSD, epilepsy, negative”, when new posts containing these descriptions, system will highlight the symptoms, generate prompt warnings and sent them to the regulatory authorities, thus help take timely response.

Figure 7 An example of automatic supervision of drug use on social media platform
Contributions

The severity of drug abuse especially among teenagers and young adults demonstrates the importance of this research. Social media has become an indispensable part for young people to share their life experience and gather information, playing an unreplaceable role in detecting and preventing non-medical drug use. However, previous research has mainly focused on platforms such as Twitter, which has limitations in inputting characters and may lead to a lack of details. More perceptions and behavior patterns need to be found through other social media. Our study collected massive posts from social forums containing large information of drug usage and analyzed drug types, causes, and symptoms across users with age characteristics. Specifically, we found that: (1) Drug topics contained 6 categories including alcoholic, tobaccos, prescriptions, hallucinogens, ecstasies, and other highly addictive drugs. Tobaccos (n=11,947), hallucinogens (n=9244), and other
**highly addictive drugs** (n=5733) were the most discussed drug categories. (2) 9646 (65% of the total posts that had age information) posts on Reddit were sent by users between 15 and 25 years old. There was relevance between age and drug addiction. People tended to change their drug usage from a primary drug (e.g. marijuana) to other highly addictive drugs (e.g. heroin) as they aged. The age groups age<15 and 15<=age<20, 20<=age<25 had very similar drug topic patterns, where the Pearson correlation coefficient (PCC) between age<15 and 15<=age<20 is .994 (P<.001), the Pearson correlation coefficient (PCC) between age<15 and 20<=age<25 is .985 (P<.001). Similar phenomena were also obtained for age groups 20<=age<25 and 25<=age<30, where the Pearson correlation coefficient (PCC) is .989 (P <.001).

(3) The reasons users take different types of drugs can be classified as *curious*, *anxious* and *pain*. The most frequently mentioned cause is curiosity about a novel experience (1215, 1215/1848), which may due to the characteristics of young users. For the curiosity cause, the ratios of posts discussing tobaccos and hallucinogens were 0.5045 (613/1215) and 0.3975 (483/1215) respectively. (4) The symptoms caused by taking drugs can be classified into 6 topics, including *hallucination*, *comfort*, *lose_control*, *affect_brain*, *anxious* and *pain*. Anxiety was the most popular symptom mentioned among the posts (574, 574/1096). The number of posts reporting anxious symptom caused by drugs is a little larger than that of posts discussing curing anxiety by drugs (n=558).
Limitations

We only collected post data from /r/drugs subreddit, which could be enlarged by adding posts from /r/addiction subreddit, /r/drug_dealer subreddit and /r/trees subreddit (a community discusses marijuana).

We used word2vec to analyze word association when extracting topics from posts. This method solved the semantic fuzzy, misspellings, slang and other issues to a certain extent. However, other deep-learning methods can be applied to analyze word association when extracting topics from posts.

We took all kinds of drug usage behavior into consideration when conducting data analysis, including medical use, misuse, and abuse. Defining the specific characteristics of drug use is needed when performing deep analysis in future work.

Future Research

We envision three possible approaches for future study. Firstly, more posts from Reddit and other social media platforms could be extended as our corpus. Through analyzing posts of different social media platforms, we believe more patterns of drug usage could be obtained.

Secondly, discovering and categorizing non-medical use of substance are extremely significant in next stage research. More advanced algorithm in the field of deep learning methods such as CNN could be applied to determine non-medical use.

Thirdly, deep analysis of specific substance such as prescription opioid,
marijuana, etc could be performed, which can help find more detailed symptoms after drug usage on Reddit.

**Conclusion**

To our knowledge, this is the first study to analyze users’ perceptions and behavior patterns concerning causes and symptoms of drug usage on Reddit. We hope the findings of this paper will help policy makers better monitor and prevent drug abuse, especially aiming at teenagers and young adults.

**Authors’ contributions**

Qiudan Li, Daniel Dajun Zeng conceived the idea for this study. Qiudan Li designed the study, conducted the data analysis and wrote the manuscript. Can Wang, Hejing Liu, Daniel Dajun Zeng contributed to the manuscript and interpretation of study findings. Can Wang and Hejing Liu contributed equally to this work. All authors read and approved the final manuscript.

**Acknowledgments**

This work was supported by the US National Institutes of Health under Grant No. 5R01DA037378-05, National Natural Science Foundation of China under Grant No. 61671450, 71621002, 71272236, The Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3.

**Conflicts of Interest**
No conflict declared.

References


