r/AskAcademia: Special Considerations in the Practice of Using Reddit for Substance Use Research

Savannah Brenneke, MPH
Meredith Meacham, PhD, MPH
Amanda Bunting, PhD
Alexandra DeLucia, MSE
Nicholas Proferes, PhD

This released version of the presentation has notes included. Notes are linked to relevant slides with a \" symbol.
No financial disclosures or conflicts of interest (MM)

Research funded by NIH T32 DA007250 and K01 DA046697

Contact:
Meredith.Meacham@ucsf.edu
Assistant Professor, Department of Psychiatry & Behavioral Sciences
University of California San Francisco
u/ProjectReddibles
Observational Substance Use Research Data
Growing share of Americans say they use YouTube; Facebook remains one of the most widely used online platforms among U.S. adults

% of U.S. adults who say they ever use ...

Since 2014-15, TikTok has arisen; Facebook usage has dropped; Instagram, Snapchat have grown

% of U.S. teens who say they ever use any of the following apps or sites

Note: Respondents who did not give an answer are not shown. Pre-2018 telephone poll data is not available for YouTube, Snapchat and WhatsApp; pre-2019 telephone poll data is not available for Reddit. Pre-2021 telephone poll data is not available for TikTok. Trend data is not available for Nextdoor.
Social Media Use in 2021.

Note: Teens refer to those ages 13 to 17. Those who did not give an answer are not shown. The 2014-15 survey did not ask about YouTube, WhatsApp, Twitch and Reddit. TikTok debuted globally in 2018.
Source: Survey conducted April 14-May 4, 2022.
Teens, Social Media and Technology 2022

PEW RESEARCH CENTER
What’s Reddit?

- Social news aggregation, content rating, discussion website founded in 2005
- 6th most visited website in the United States, 10th most visited worldwide (2023)
- Content-specific communities or forums called “subreddits”
- Pseudo-anonymous contributors and volunteer moderators
- Customizable interaction and engagement
  - subscribing to subreddits
  - commenting (in nested threads)
  - upvotes/downvotes ("karma points") → visibility
How I use Reddit

Australia
r/australia

bicycling
r/bicycling

Cats
r/cats

DataIsBeautiful
r/dataisbeautiful

Drug and Alcohol Dependence
Volume 188, 1 July 2018, Pages 364-369

Full length article
Understanding emerging forms of cannabis use through an online cannabis community: An analysis of relative post volume and subjective highness ratings
Meredith C. Meacham, Michael J. Paul, Danielle E. Ramo

PLOS ONE
OPEN ACCESS PEER REVIEWED RESEARCH ARTICLE
“I got a bunch of weed to help me through the withdrawals”: Naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums
Meredith C. Meacham, Alicia L. Nobles, D. Andrew Tompkins, Johannes Thru
Published: February 8, 2022 • https://doi.org/10.1371/journal.pone.0263683
<table>
<thead>
<tr>
<th>Benefits of using Reddit</th>
<th>Challenges of using Reddit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content specific communities</td>
<td>Large amounts of data</td>
</tr>
<tr>
<td>Open discussion &amp; rich language</td>
<td>Ephemeral data</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Lack of consistent demographic &amp; geographic information</td>
</tr>
</tbody>
</table>
### Who uses Reddit in the United States? (compared to other social media)

<table>
<thead>
<tr>
<th></th>
<th>YouTube</th>
<th>Facebook</th>
<th>Instagram</th>
<th>Pinterest</th>
<th>LinkedIn</th>
<th>Snapchat</th>
<th>Twitter</th>
<th>WhatsApp</th>
<th>TikTok</th>
<th>Reddit</th>
<th>Nextdoor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>81</td>
<td>69</td>
<td>40</td>
<td>31</td>
<td>28</td>
<td>25</td>
<td>23</td>
<td>23</td>
<td>21</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td><strong>Men</strong></td>
<td>82</td>
<td>61</td>
<td>36</td>
<td>16</td>
<td>31</td>
<td>22</td>
<td>25</td>
<td>26</td>
<td>17</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td>80</td>
<td>77</td>
<td>44</td>
<td>46</td>
<td>26</td>
<td>28</td>
<td>22</td>
<td>21</td>
<td>24</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td><strong>White</strong></td>
<td>79</td>
<td>67</td>
<td>35</td>
<td>34</td>
<td>29</td>
<td>23</td>
<td>22</td>
<td>16</td>
<td>18</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td><strong>Black</strong></td>
<td>84</td>
<td>74</td>
<td>49</td>
<td>35</td>
<td>27</td>
<td>26</td>
<td>29</td>
<td>23</td>
<td>30</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td><strong>Hispanic</strong></td>
<td>85</td>
<td>72</td>
<td>52</td>
<td>18</td>
<td>19</td>
<td>31</td>
<td>23</td>
<td>46</td>
<td>31</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td><strong>Ages 18-29</strong></td>
<td>95</td>
<td>70</td>
<td>71</td>
<td>32</td>
<td>30</td>
<td>65</td>
<td>42</td>
<td>24</td>
<td>48</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>30-49</td>
<td>91</td>
<td>77</td>
<td>48</td>
<td>34</td>
<td>36</td>
<td>24</td>
<td>27</td>
<td>30</td>
<td>22</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>50-64</td>
<td>83</td>
<td>73</td>
<td>29</td>
<td>33</td>
<td>33</td>
<td>12</td>
<td>18</td>
<td>23</td>
<td>14</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>65+</td>
<td>49</td>
<td>50</td>
<td>13</td>
<td>18</td>
<td>11</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

Pew Research Center 2021
Changes and Uncertainties

**TECHNOLOGY**

Thousands of Reddit communities ‘go dark’ in protest of new developer fees  
June 12, 2023 - 5:00 AM ET  
By Bobby Allen, Tima Wilson

**COALITION for INDEPENDENT TECHNOLOGY RESEARCH**

Signatures & Mutual Aid - Restricting Reddit Data Access Threatens Online Safety & Public-Interest Research

Last week, soon after Reddit announced plans to restrict free access to the Reddit API, the company cut off access to Pushshift, a data resource widely used by communities, journalists, and thousands of academics worldwide. The impacts of losing access to Reddit data are twofold:

- It will disrupt critical projects from thousands of journalists, academics, and civil society actors who study some of the most important issues impacting our societies today, and
- It will hinder Reddit's volunteer moderators from the vital positions they hold keeping their communities and Reddit's user base safe from harm.
Workshop Overview

Activity 1: Exploring Reddit
Rapid Use of Reddit to Understand Emerging Drug Trends
Amanda Bunting, PhD
New York University

Activity 2: Reddit Research Questions
r/AskAComputerScientist: Processing Reddit Data for the Social Sciences
Alexandra DeLucia, MSE
Johns Hopkins University

Activity 3: Working with Reddit Data
Harnessing Reddit: The Methods and Their Limitations in Analyzing Unstructured Data from Social Media
Savannah Brenneke, MPH
Johns Hopkins University

Ethical Responsibility and Reddit Research: How Contextual Integrity Can Help Guide Practice
Nicholas Proferes, PhD
Arizona State University

Q & A | Discussion | Further Resources
Activity 1: Exploring Reddit

Go to Reddit.com or download the Reddit App. Search for a topic or community related to your research interests:

Or explore communities by topic:

Pick a subreddit community and note:

- What stands out?
- How is the community described? When did it start?
- How many subscribers does it have?
- What kind of posts are there (text, links, image, video)?
- What are the subreddit rules?
Activity 1 Example: r/askdrugs

r/askdrugs Rules

1. Ensure information is accurate and reduces harm
2. Don't discuss places to buy or sell (sourcing)
3. Don't be judgemental, be positive and constructive
4. No drug identification
5. Drug test questions are allowed

Strictly no requesting, mentioning or giving sources of drugs or paraphernalia, whether legal or illegal. If in doubt - if your post, or a reply to it would make it easier for someone to get drugs, it’s not permitted. This includes sourcing conducted in private messages.

This does not include harm reduction related paraphernalia** such as scales, testing kits, syringes, micron filters and so on.
RAPID USE OF REDDIT TO UNDERSTAND EMERGING DRUG TRENDS

Amanda Bunting, PhD
@ABuntingPhD
Acknowledgments

• No conflicts of interest to report

• Funding: K01DA053435, R25DA037190

• Collaborators
  – Noa Krawczyk
  – David Frank
  – Marie Bragg & Lab
  – Sam Friedman
  – Yuanqi Gu
  – Thomas Lippincott
  – Meredith Meacham
  – Simran Arya
  – Suhas Nagappala
Use of Reddit to examine opioid use during COVID-19 lockdowns

Socially-supportive norms and mutual aid of people who use opioids: An analysis of Reddit during the initial COVID-19 pandemic

Amanda M. Bunting, David Frank, Joshua Arshonsky, Marie A. Bragg, Samuel R. Friedman, Noa Krawczyk

Informal Coping Strategies Among People Who Use Opioids During COVID-19: Thematic Analysis of Reddit Forums

Josh Arshonsky, Noa Krawczyk, Amanda M Bunting, David Frank, Samuel R Friedman, Marie A Bragg

COVID-19–Related Changes to Drug-Selling Networks and Their Effects on People Who Use Illicit Opioids

Journal of Studies on Alcohol and Drugs, 84(2), 222–229 (2023).

David Frank, Noa Krawczyk, Joshua Arshonsky, Marie A. Bragg, Samuel R. Friedman, + Affiliations
Fentanyl and other synthetic opioid overdose involved fatalities have been rising since 2013.
Use of Reddit to understand how persons who drugs were experiencing the rise of fentanyl in the drug supply

Fentanyl in Pressed Oxycodone Pills: A Qualitative Analysis of Online Community Experiences with an Emerging Drug Trend

Simran Arya, Suhas Nagappala, Noa Krawczyk, Yuanqi Gu, Meredith C. Meacham & Amanda M. Bunting

Trends in fentanyl content on Reddit substance use forums, 2013-2021

Accepted Journal of General Internal Medicine
Is your research question appropriate for Reddit data?

• Accelerated data collection
  – Overcomes some of the limitations of the scientific process, data collection

• Real time
  – Data can be collected historically or current

• Subpopulation behaviors that are organic
  – Less likely to have social desirability effects, stigma reduced
  – Explore phenomenon that we are ethically unable to explore

• Resources to investigate are otherwise limited
  – Only if all the above are still important
Example research questions

• What are the motives of substance use? (e.g., Pestana et al., 2021; Sharma et al., 2017)

• How has substance use (and characteristics of use) changed over time? (e.g., Bowen et al., 2019; Bunting et al., 2023; Meacham et al., 2018)

• How do persons who drugs resonate with televised portrayals of substance use? (e.g., Kaufman et al., 2020)

• How do persons who drugs support each other? (e.g., D’Agostino et al., 2017; Bunting et al., 2021; Rhidenour et al., 2021)
National Drug Early Warning System

Web Surveillance

Social media platforms provide a unique indicator of activity in the rapidly changing market for new psychoactive substances (NPS). On the popular Reddit website, drug-related discussion occurs in user-generated subreddits dedicated to a drug, a class of drugs, or more general topics related to drug use or experimentation. The contents of drug subreddits can be useful for estimating temporal trends in NPS use, including early and real-time identification of emerging drugs.

In collaboration with the Machine Perception and Cognitive Robotics (MPCR) Lab, under the direction of Dr. Barenholz, PhD candidate Paul Morris, and PhD student Daniel Van Zant the NDEWS Coordinating Center developed a web monitoring platform for early detection of NPS in drug subreddits. Trends in drug discussion are quantified through anonymized, aggregate keyword counts derived from algorithmic monitoring of ~60 active drug-oriented subreddits. Keyword metrics count mentions over time of keywords that refer to a drug. Machine learning models are employed for automated detection and aggregation of drug keywords.

In combination, these methods detect drug-related activity that is anomalous and potentially indicative of an emerging trend in the development or use of novel substances. Validation on historical trends reveals the detection of an NPS in drug subreddit discussion is predictive of its subsequent emergence in toxicology and other real-world signals. Early detection of NPS trends by web monitoring serves as a source for further investigation and collaboration with NDEWS partners.

https://ndews.org/novel-surveillance/web-surveillance/
Activity 2: Reddit Research Questions

- Do you want to focus on a single community or many?
- What time periods are you interested in?
- Are your interests qualitative, quantitative, or both?
- What has been published on this topic (biomedical, social sciences, computer sciences)?
r/AskAComputerScientist: Processing Reddit Data for the Social Sciences

Alexandra DeLucia, MSE
PhD Candidate
Center for Language and Speech Processing
Department of Computer Science, Johns Hopkins University

Please direct all questions or comments to aadelucia@jhu.edu or @alexir563
## Overview

<table>
<thead>
<tr>
<th>1</th>
<th>Tenets of Data Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Anatomy of a Reddit Thread</td>
</tr>
<tr>
<td>3</td>
<td>Processing Reddit Data</td>
</tr>
<tr>
<td>3.1</td>
<td>Obtaining + Future of APIs</td>
</tr>
<tr>
<td>3.2</td>
<td>Storing</td>
</tr>
<tr>
<td>3.3</td>
<td>Cleaning</td>
</tr>
<tr>
<td>4</td>
<td>TL;DR + AMA</td>
</tr>
</tbody>
</table>
No conflicts of interest to report
Anatomy of a Reddit Thread

- Submission interactions
  - Subreddit name
  - Posting user
  - Date created
  - Title
  - Content
  - Number of comments
  - Upvotes
  - Awards
  - Commenting user
  - Date created
  - Comments / replies
  - Upvotes
  - Search comments
Anatomy of a Reddit Thread...in code*

```json
{
    'id': '14ludhx',
    'permalink': '/r/opiates/comments/14ludhx/what_is_a_high_dose_of_oxycodeone/',
    'author': Redditor(name='spatsink'),
    'author_flair_text': 'Fresh Account',
    'title': 'what is a high dose of oxycodone?',
    'selftext': '',
    'is_self': True,
    'url': 'https://www.reddit.com/r/opiates/comments/14ludhx/what_is_a_high_dose_of_oxycodeone',
    'score': 21,
    'ups': 21,
    'upvote_ratio': 0.89,
    'num_comments': 93,
    'num_crossposts': 0,
    'num_duplicates': 0,
    'num_reports': None,
    'comments': [],
    'over_18': False,
    'edited': False,
    'gilded': 0,
    'gildings': {},
    ..
}  
```

*This is a JavaScript Object Notation (JSON) object, it is a way to store structured data. Similar to comma-separated values (CSV).
Processing Reddit Data
Obtaining Reddit Data

Two main methods for obtaining Reddit data

1. Application Programming Interfaces (APIs)*
   a. Pushshift.io
   b. Official Reddit API

2. Manually (via Reddit Search bar)
   a. Screenshots
   b. Copy/paste

*API access is currently undergoing changes due to Reddit policy
Obtaining Reddit Data

Reddit API

• Allows for search, collecting specific posts and comments, collecting subreddit-specific posts
• Contains up-to-date post attributes (i.e., upvotes)

Pushshift.io

• Third-party (i.e., unaffiliated) collection and host of Reddit data
• Large "dumps" available containing all comments/submissions for all subreddits per month (roughly 30GB per compressed file)
• Search for subreddits, keywords, and specific date ranges, and other archived posts
Future of Reddit API

- Official Reddit API Access for researchers and third-party apps is changing this year. User access remains the same at no cost.
- Effective July 1, the free tier will allow 100 queries (requests) per minute.
- Pushshift.io is currently down will potentially allow access for moderators and researchers.
- Adult-only (NSFW) subreddits and posts could be restricted, which includes some substance-use communities.
Storing the Data

In a Database

- MySQL, Apache Solr
- Re-running analysis with various settings is easier (e.g., specific date ranges, excluding users)
- Learning curve for storing and querying

In Files

- CSVs, JSON, etc
- Requires programming knowledge for efficient reading
- Iterate through all files for each analysis

How to Store?
Storing the Data

- Store *compressed* files (e.g., .gz)
  - Text is significantly smaller than images and videos

- Only save the needed information
  - Not all attributes are needed (e.g., "author flair")
  - Thorough cleaning can reduce storage needs

- Save files grouped in ways they will be frequently accessed
  - e.g., by subreddit, by date, by post (grouped with comments)
Cleaning the Data

1 Filtering unwanted data
   • Spam, duplicates, bots, ads
   • Restricting to posts with specific keywords
   • Removing posts with images, crossposts

2 Post-level processing
   • Removing HTML and Markdown formatting code
   • Stem words for frequency stats
     • e.g., {smoking, smoked, smoke} -> smok
TL;DR

Please direct all questions or comments to aadelucia@jhu.edu or @alexir563
Harnessing Reddit: The Methods and Their Limitations in Analyzing Unstructured Data from Social Media

Savannah G. Brenneke, MPH
sbrenne7@jh.edu | @sbrennekek
Bloomberg School of Public Health
Johns Hopkins University
Disclosures
I have NO financial disclosure or conflicts of interest with the presented material

Funding
NIDA T-32 Drug Dependence Epidemiology Training Program T32DA007292
Overview

• Assumptions in Reddit Data:
  • Representativeness
  • Generalizability

• Quantitative Methods
  • Computational
  • Infodemiologic

• Qualitative Methods
  • Focus groups & interviews
  • Observation

• Example – Meacham, M.C., Nobles, A.L., Tompkins, D.A., & Thrul, J. (2022)
Assumptions within Reddit
WHO is using Reddit

- More than 430 million users MONTHLY
- >60% are between 18-29 years of age
- Majority of users are:
  - MALE
  - LEFT-LEANING
  - COLLEGE EDUCATED
- Over 90% of posts are in ENGLISH *including non-native speakers
Estimating Demographic Information

- Individuals OCCASSIONALLY disclose demographic details

- Nature of disclosures vary, with disclosures serving certain purposes
  - Narration
  - Establish credibility
  - Relate to topic information

- Researchers, e.g. Chew et al (2021), have attempted to use computational methods to estimate user demographics
What people are DOING on Reddit

Assumptions

Quantitative

Qualitative

Example

About Community

A community for images, videos, discussions, artwork, and everything dachshund related. Feel free to share your doxie!

Created Dec 6, 2008

341k Members

716 Online

Top 1%

Ranked by Size

/r/Dachshund Rules

1. All content must be related to dachshunds
2. Keep the community wholesome
3. No Impersonation
4. Follow reddiquette, don’t ask for upvotes
5. Self-promotion is only allowed for active community members
6. Help by reporting spam

Special needs dogs seem even more loving than usual. I adopted a blind and deaf dachshund and she was amazing. She still follows me around 8:30 in the morning, and she would steal the end of the toilet paper roll and run it out to the living room and chew it up. My other dog got worried until she was caught.

No questions but she looks like a beautiful sweet boy

Aw, thank you! He was definitely a beautiful boy.

Ours is fully blind and has lost half hearing now. Did yours adjust okay to not seeing/hearing?
Redditor Engagement

**Assumptions**

- Quantitative
- Qualitative
- Example

**POSTER**

**COMMENTER**

**LURKER**

me

hmm... a/ hasn't posted anything
User Level Active-Engagement: SUBMISSIONS

- Redditor is initiator and creator of original content shared in community
- Initiation can be high-stakes for some
- Submissions as data points may be informative for:
  - Story-telling / Narration
  - Advice-seeking
  - General disclosure
  - Entertainment/Trend Sharing
User Level Active-Engagement: COMMENTING & INTERACTING

- Multiple levels of engaging without being an "initiator":
  - Commenting
  - Up & Down Voting
  - Gifting
  - Cross posting

- Analyses may examine submissions + comments, comments only, other engagement like Karma and Voting
  - See metadata
  - Analysis of submissions + comments inform discourse, community, support networks, etc.

- Special considerations for topic drift*
User Level Passive-Engagement: ‘LURKING’

- Passively consuming information with OR without an account
- Browsing
- Searching

- 'lurkers' or passive-consumers cannot be captured in the data

- Belonging to and content consumption behaviors could be insightful to current or future 'real-life' behaviors

Assumptions

Quantitative

Qualitative

Example
## Why does all this matter?

<table>
<thead>
<tr>
<th>Representativeness</th>
<th>Generalizability</th>
</tr>
</thead>
</table>
| - Reddit user base is unlikely to be representative of populations  
  - Minority  
  - Disconnected  
  - Lower-income  
  - Non-native English speakers  
  - Engagement levels have different risks & rewards, i.e. experiences shared/discussed may skew data | - Without demographic and geographic information, Reddit data is unlikely to generalize to broader populations  
  - Across Reddit generalizability is also limited by  
    - Community rules  
    - Membership  
    - Reward systems |
QUANTITATIVE METHODS
Computational Methods

- **Term frequency – Inverse document frequency (tf-idf)**
  - Estimation of the most commonly used unique and informative terms within a document and among a collection of documents
  
  \[ \text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t) \]
  
  \[ \text{idf}(t) = \log \frac{1+n}{1+df(t)} + 1 \]

- **Word-embeddings**
  - Numerical (vector) representation of a word that allows words that occur in similar contexts to have the similar representations.
  - E.g., word2vec, GloVe, BERT, ClinicalBERT

- **Sentiment analysis**
  - Context mining of subjective information
  - Text classification (e.g., positive, negative, neutral)
Computational Methods

- **Topic modeling**
  - Exploring abstract “topics” within a set of documents
  - Latent Dirichlet allocation (LDA)

- **Network analysis**
  - Structural analysis of relationships or processes

- **Predictive modeling**
  - Using existing data to forecast future
  - Regression models, decision trees and neural networks

**Assumptions**
- Quantitative
- Qualitative

**Example**
Infodemiologic Methods

- Infodemiology coined in early 2000’s by Dr. Gunther Eysenbach
- “…the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy.”
- Supply- and demand-based infodemiology
  - Supply = what is published online
  - Demand = search and click behaviors
- Infoveillance as a second arm of infodemiology
SUPPLY: Information (Concept) Prevalence, Incidence, Co-occurrence

- **Prevalence** —
  - Absolute or relative number a term or keyword is present in a pool of information, i.e. Reddit submissions, comments or submissions + comments

- **Incidence** —
  - Number of new concept units per unit of time, i.e. incidence of mentions of

- **Co-occurrence** —
  - Number of two or more concepts present in the same unit, i.e. two concepts mentioned in a given Reddit submission
Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants

Alicia L. Nobles, Eric C. Leas, Theodore L. Caput, Shu-Hong Zhu, Steffenie A. Strathdee & John W. Ayers

Evaluation of Artificial Intelligence Responses to Public Health Questions

John W. Ayers, PhD, MA(2), Zeceliah Zhu, BS(3), Adam Poljak, PhD(3), et al.

Characterizing Help-Seeking Searches for Substance Use Treatment From Google Trends and Assessing Their Use for Involvement: Longitudinal Descriptive and Validation Statistical Analysis

Patton 1, Daniel Abravitz 2, Johnson D. Leas 3, Nobles 3, Caput 1, Ayers 1, Th Brenda 1, Börjeson 1


Mike Conway, Cole Ciferri, and Annie T. Chen

Assumptions

Quantitative

Qualitative

Example
QUALITATIVE METHODS
Observational Qualitative Methods

- **CONTENT ANALYSIS**: most commonly used qualitative method – assuming treated as a qualitative
  - Note: quant + qual approaches to content analysis are common

- Require development of a **codebook** to guide coding of content
  - *Inductive* – ground-up approach
  - *Deductive* – top-down approach
  - A mix of inductive-deductive

- Others:
  - Discourse analysis
  - Ethnographic observation
  - Grounded theory
  - Case Studies
"I got a bunch of weed to help me through the withdrawals": Naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums

Meredith C. Measham, Alicia L. Nobles, D. Andrew Tompiński, Johannes Throl

Published: February 8, 2022 • https://doi.org/10.1371/journal.pone.0263583

Abstract

A growing body of research has reported on the potential opioid-sparing effects of cannabis and cannabinoids, but less is known about specific mechanisms. The present research examines cannabis-related posts in two large online communities on the Reddit platform ("subreddits") to compare mentions of naturalistic cannabis use by persons self-identifying as actively using opioids versus persons in recovery. We extracted all posts mentioning cannabis-related keywords (e.g., "weed", "cannabis", "marijuana") from December 2015 through August 2019 from an opioid use subreddit and an opioid recovery subreddit. To investigate how cannabis is discussed at-scale, we identified and compared the most frequent phrases in cannabis-related posts in each subreddit using term-frequency-inverse document frequency (TF-IDF) weighting. To contextualize these findings, we also conducted a qualitative content analysis of 200 random posts (100 from each subreddit). Cannabis-related posts were about twice as prevalent in the recovery subreddit (n = 906; 5.4% of 16,791 posts) than in the active opioid use subreddit (n = 4,224; 2.6% of 159,994 posts, p < .001). The most frequent phrases from the recovery subreddit referred to time without using opioids and the possibility of using cannabis as a "treatment." The most frequent phrases from the opioid subreddit referred to concurrent use of cannabis and opioids. The most common motivations for using cannabis were to manage opioid withdrawal symptoms in the recovery subreddit, often in conjunction with anti-anxiety and GI-distress "comfort meds," and to enhance the "high" when used in combination with opioids in the opioid subreddit. Despite limitations in generalizability from pseudonymous online posts, this examination of reports of naturalistic cannabis use in relation to opioid use identified withdrawal symptom management as a common motivation. Future research is warranted with more structured assessments that examines the role of cannabis and cannabinoids in addressing both somatic and affective symptoms of opioid withdrawal.
Concept prevalence, reported as “proportion of cannabis-related posts”:
- Cannabis-related posts overall by subreddit and estimations overtime (per month)

tf-idf, examining most frequent words and phrases in cannabis-related posts
- Explored top 10 unigram, bigrams and trigrams, via highest tf-idf weighting

Assumptions

Quantitative

Table 1. Comparison of posts containing cannabis terms between active opioid use and opioid recovery subreddits (late 2015-mid 2019).

<table>
<thead>
<tr>
<th></th>
<th>Recovery Subreddit</th>
<th>Opioid Subreddit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Posts (N)</td>
<td>16,791</td>
<td>159,994</td>
</tr>
<tr>
<td>Cannabis Posts (N, % of Total)</td>
<td>908 (5.4%)</td>
<td>4,224 (2.6%)</td>
</tr>
<tr>
<td>Total Wordcount (Median, IQR)</td>
<td>101 (32, 124)</td>
<td>24 (8, 88)</td>
</tr>
<tr>
<td>Cannabis Post Wordcount (Median, IQR)</td>
<td>279 (151, 435)</td>
<td>173 (80, 347)</td>
</tr>
<tr>
<td>Change in proportion of posts mentioning cannabis per month (Slope, p-value)</td>
<td>0.014% 0.29</td>
<td>-0.013% 0.0051</td>
</tr>
</tbody>
</table>

IQR: Inter-quartile range.

Table 2. Top 10 most frequent n-grams in cannabis-related posts

<table>
<thead>
<tr>
<th></th>
<th>Recovery unigrams</th>
<th>Opioid unigrams</th>
<th>Recovery bigrams</th>
<th>Opioid bigrams</th>
<th>Recovery trigrams</th>
<th>Opioid trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marijuana, cbd, clean, opiate, recovery, cannabis, kratom, day, days, weed</td>
<td>opiates, marijuana, cbd, heroin, weed, like, smoke, oxy, just, get</td>
<td>day marijuana, weed recovery, zoloft pot, health benefits, medical marijuana, marijuana treat, opiate substitution, substitution thc, benefits cannabis, weed social</td>
<td>smoke weed, smoking weed, weed opiates, first time, medical marijuana, hydro weed, weed hydro, feel like, anyone else, right now</td>
<td>pill porn weed, coke h weed, combining tramadol weed, going weed withdrawals, weed valium amazing, xanax ambien weed, growing cannabis mistake, bars weed hash, h alc weed, 10mg hydro weed</td>
<td>opiates, marijuana, cbd, heroin, weed, like, smoke, oxy, just, get</td>
<td>smoke weed, smoking weed, weed opiates, first time, medical marijuana, hydro weed, weed hydro, feel like, anyone else, right now</td>
</tr>
</tbody>
</table>
Quantitative Methods: “Keyness” scoring

Keyness scoring, estimating terms or words that are unique and important within a subreddit by comparing to the other

Recovery subreddit key terms focus on
- Time spent "clean"
- Withdrawal & PAWS
- Relapse
- Challenges with recovery

Larger opioid subreddit key terms
- Mentions of people
- Profanity
- Tolerance
- References of amounts or effects of drugs

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 15 most unique unigrams and bigrams in cannabis-related posts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top 15 most unique terms (unweighted)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recovery unigrams</strong></td>
</tr>
<tr>
<td><strong>Recovery bigrams</strong></td>
</tr>
<tr>
<td><strong>Opioid unigrams</strong></td>
</tr>
<tr>
<td><strong>Opioid bigrams</strong></td>
</tr>
</tbody>
</table>
Qualitative Methods: Content analysis

<table>
<thead>
<tr>
<th>Code Definition</th>
<th>Recovery Subreddit (N = 100)</th>
<th>Opioid Subreddit (N = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positive</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Opioid use disclosure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No / Not personal / NA</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>Yes</td>
<td>89</td>
<td>78</td>
</tr>
<tr>
<td>Past</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Current</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>Reducing</td>
<td>74</td>
<td>29</td>
</tr>
<tr>
<td><strong>Cannabis use disclosure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No / Not personal / NA</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>Yes</td>
<td>77</td>
<td>68</td>
</tr>
<tr>
<td>Past</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Current</td>
<td>52</td>
<td>49</td>
</tr>
<tr>
<td>Reducing</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td><strong>CBD use disclosure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specifically mentions CBD use</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td><strong>Motivations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Withdrawal management</td>
<td>43</td>
<td>12</td>
</tr>
<tr>
<td>Helpful on its own</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Helpful with other strategies</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Not helpful</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Questioning compatibility</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Polysubstance use</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Pain management</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Helpful with other strategies</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Not helpful</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Questioning compatibility</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polysubstance use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain management</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare provider interaction</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>Related to cannabis use</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

**Sample Quotations**

**Withdrawal symptom management**
"Marijuana is AMAZINGLY helpful. I'm going to ramble because it is a nice distraction. I hate marijuana usually... I had some pretty bad WDs and I feel better than ever... because of weed." (recovery subreddit)

"2 days clean. I have no one else to tell. ... I'm taking a few hits of pot every 4 hours to ease the symptoms... it's been making it tolerable." (recovery subreddit)

**Questioning compatibility**
"As far as marijuana goes, I haven't decided if I want to start again. Any feedback from people who have stopped marijuana and then started again would be appreciated. Any regrets starting again? Did it affect your recovery negatively?" (recovery subreddit)

"Opiates are usually my first priority, but I'm interested in weed now. My options are either maintenance therapy without weed or to try to medicate with weed. ... Should I keep pursuing maintenance? Or really try to get clean and smoke weed instead? Maybe it's just that now that I'm on sub i see maintenance does not mean I am high all the time and I still want to keep doing that somehow." (recovery subreddit)

**Polysubstance use**
"I'm having a very pleasant night sitting here, vegging out online—I am currently stable on my Suboxone dose, with a nice shot of crystal and some good weed..." (opioid subreddit)

"Tonight I smoked up a joint of heady weed with hash in it, then smoked like 500mg heroin 50mg at a time." (opioid subreddit)

**Pain management**
"Hi chronic pain sufferers—how effective is weed for your pain? ... does it work for you at all? I find that it is just a distraction for me when I get random back pain." (opioid subreddit)

"I came off methadone 9 days ago, and the withdrawal symptoms have sucked, sneezing, watery eyes, restless legs. You know, all that fun stuff. I've been using medicinal cannabis to control the pain, but I'm getting really fed up. Has anyone else done this, and if so, about how long does methadone withdrawal usually last?" (recovery subreddit)

**Direct quotations are lightly edited to reduce chances of re-identification.**
Activity 3: Working with Reddit Data

• What information do you want to access, store, and analyze?

• What data might be missing (as it relates your research question)?

• What challenges might you encounter?

• Who could you collaborate with in computer science, data science, information science for data collection (if needed)?

• Could you collaborate with in clinical, health services, policy, civil society, or other fields?
Ethical Responsibility and Reddit Research: How Contextual Integrity Can Help Guide Practice

Nicholas Proferes
Arizona State University
nprofere@asu.edu
DISCLOSURES

No conflicts of interest to report
Talk Outline

I. Ethics: Two Ways

II. Gap between what researchers do and what users expect.

III. Contextual Integrity
Ethics as *Contemplation*

**Ethics:** systems of principles that we use to guide us in making moral evaluations.
- Utilitarianism
- Deontology
- Virtue Ethics, etc.

Using our capacities for **reason** and **judgment** to critically examine our actions and our character.
“Ethics” as *Compliance*

Process of ensuring conformity with all relevant laws, policies, and guidelines.

These laws/guidelines/policies often draw on ethics-based principles, such as:

- Respecting human dignity and autonomy
- Maximizing benefit/minimizing harm
- Justice and beneficence

Ensuring researchers do not violate certain baseline conditions for the treatment of others.
“Public” social media data analysis often not considered (by U.S. IRBs) as “research involving human subjects.”

But, projects using publicly available data from social media sites do not lack ethical dimensions.

They require contemplation alongside (often minimal) compliance.
“Do users know their content is being used for academic study and how do they feel about it?”

Surveyed Twitter users, asking:

Whether users think researchers are “allowed” to use their content without re-consent.

Levels of comfort with their tweets being used.

Contextual factors (permission asked; study content; size of dataset; who is doing analysis; kinds of content; status of content; quoted in study).
<table>
<thead>
<tr>
<th><strong>Question</strong></th>
<th><strong>Very uncomfortable</strong></th>
<th><strong>Somewhat uncomfortable</strong></th>
<th><strong>Neither uncomfortable nor comfortable</strong></th>
<th><strong>Somewhat comfortable</strong></th>
<th><strong>Very comfortable</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>... you were not informed at all?</td>
<td>35.1%</td>
<td>31.7%</td>
<td>16.4%</td>
<td>13.4%</td>
<td>3.4%</td>
</tr>
<tr>
<td>... you were informed about the use after the fact?</td>
<td>21.3%</td>
<td>29.1%</td>
<td>20.5%</td>
<td>22.0%</td>
<td>7.1%</td>
</tr>
<tr>
<td>... it was analyzed along with millions of other tweets?</td>
<td>2.6%</td>
<td>18.7%</td>
<td>25.5%</td>
<td>30.0%</td>
<td>23.2%</td>
</tr>
<tr>
<td>... it was analyzed along with only a few dozen tweets?</td>
<td>16.5%</td>
<td>30.3%</td>
<td>24.0%</td>
<td>20.2%</td>
<td>9.0%</td>
</tr>
<tr>
<td>... it was from your “protected” account?</td>
<td>54.9%</td>
<td>20.5%</td>
<td>13.8%</td>
<td>6.0%</td>
<td>4.9%</td>
</tr>
<tr>
<td>... it was a public tweet you had later deleted?</td>
<td>31.3%</td>
<td>32.5%</td>
<td>20.5%</td>
<td>10.4%</td>
<td>5.2%</td>
</tr>
<tr>
<td>... no human researchers read it, but it was analyzed by a computer program?</td>
<td>2.6%</td>
<td>14.3%</td>
<td>30.5%</td>
<td>32.3%</td>
<td>20.3%</td>
</tr>
<tr>
<td>... the human researchers read your tweet to analyze it?</td>
<td>9.7%</td>
<td>27.6%</td>
<td>25.0%</td>
<td>25.4%</td>
<td>12.3%</td>
</tr>
<tr>
<td>... the researchers also analyzed your public profile information, such as location and username?</td>
<td>32.2%</td>
<td>23.2%</td>
<td>21.0%</td>
<td>13.9%</td>
<td>9.7%</td>
</tr>
<tr>
<td>... the researchers did not have any of your additional profile information?</td>
<td>4.9%</td>
<td>15.4%</td>
<td>25.1%</td>
<td>34.1%</td>
<td>20.6%</td>
</tr>
<tr>
<td>... your tweet was quoted in a published research paper, attributed to your Twitter handle?</td>
<td>34.3%</td>
<td>21.6%</td>
<td>21.6%</td>
<td>13.1%</td>
<td>9.3%</td>
</tr>
<tr>
<td>... your tweet was quoted in a published research paper, attributed anonymously?</td>
<td>9.0%</td>
<td>16.8%</td>
<td>26.5%</td>
<td>28.4%</td>
<td>19.4%</td>
</tr>
</tbody>
</table>
User understandings of data-uses are limited.

Important to understand and study phenomena.

Levels of acceptance vary by specifics.

Notifying/consenting users could create anxiety.
Nissenbaum’s framework of privacy as contextual integrity helps us unpack possible issues.

Every situation in life has:

• **Norms of Information Appropriateness**
  At a health check-up, it’s appropriate for a doctor to ask about my health conditions, but probably not my bank balance.

• **Norms of Information Flow**
  At a health check-up, it’s appropriate for my doctor to share my health information about with my insurance, but probably not my bank.
Where do these normative expectations come from?

For each contextual situation, we consider:

• Social roles (ours and others)

• Expectations we have for others, and that we think others have of us.

• Likely actions and practices of others, as informed by history, culture, law, and social convention.
Addressing tensions…

• We must continuously weigh the tension between our values and duties.

• Try to get a sense of the norms of appropriateness and information flow with the subreddit.
  • Size, discoverability, subreddit rules.

• In data-sharing, reporting, find ways to anonymize user-data when possible.
  • Obfuscate quotes, don’t include usernames, potentially subreddit names, images.
Addressing tensions…

• Re-sharing by request only, with your own user agreement.

• Document decision-making and discuss it!

• Ethical obligations to others: co-workers and students.
  • Debriefing as a team
  • Support services
• Association of Internet Researchers.

• Just shy of two decades of tackling ethics questions involving Internet Research.

• Has published two ethical decision-making guides + flowchart: https://aoir.org/ethics/

• 2019 Ethical Guidelines (v 3.0).
Extra Information
Anonymization

Reddit is a pseudo-anonymous website but there are further measures that can be taken to protect identity

• Replace each unique username with unique, different ID (e.g., "User1"..."User234"
Reddit API Resources

• JSON example https://www.json.org/json-en.html
• Official API https://www.reddit.com/dev/api/
• Submission and comment attributes https://docs.google.com/document/d/1xpOD3dh9CgCrrRP TBsEfObdgYNstHQU6M1DXrLgUrCc/edit?usp=sharing
Changes to Reddit API Access

"Though access to Pushshift data for research purposes is not available at this time, we are keen to explore possibilities that might allow us to provide researchers with access to datasets essential for their valuable social media research."

https://www.reddit.com/r/pushshift/comments/13w6j20/advancing_communityled_moderation_an_update_on/

"Effective July 1, 2023, the rate limits to use the Data API free of charge are.. 100 queries per minute per OAuth client id if you are using OAuth authentication and 10 queries per minute if you are not using OAuth authentication."

https://www.reddit.com/r/reddit/comments/145bram/addressing_the_community_about_changes_to_our_api/
Manual Reddit Search

- Search syntax [https://www.reddit.com/wiki/search/](https://www.reddit.com/wiki/search/)
## Data Size Estimates

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single comment or submission</td>
<td>~1KB</td>
</tr>
<tr>
<td>All submissions and comments from r/trees (uncompressed)</td>
<td>3GB</td>
</tr>
<tr>
<td>Month of all comments (compressed)</td>
<td>30GB</td>
</tr>
</tbody>
</table>

Table only includes text and metadata information, not images or videos
Apache Solr for Reddit

• Getting started with Solr https://solr.apache.org/guide/solr/latest/index.html
• Solr configuration for Reddit https://github.com/AADeLucia/Solr-Reddit-Utils
Post-Level Processing

General great resource for tutorials from The Summer Institutes in Computational Social Science [https://sicss.io/overview](https://sicss.io/overview)

- **Python resources**
  - Removing unwanted symbols in text can be done with Python regular expressions ("regex", or pattern-matching) and/or NLTK
  - Scikit-learn [https://scikit-learn.org/stable/](https://scikit-learn.org/stable/)
  - Loading the files and removing duplicates can be done with Pandas. The `read_json` method is especially useful.

- **R resources**
  - LOL learn Python
  - quanteda package for analyzing text data [http://quanteda.io](http://quanteda.io)
  - But please, learn Python.
Topic Drift

• Jerry Hobbs original overview of “topic drift” in natural human-to-human conversations:
  • https://www.isi.edu/~hobbs/discourse-references/topic-drift.pdf

• ResearchGate Article on the quality issues faced in social media date:
  • Nitin Agrawal, January 2010

• JMIIR Publication of a study exploring topic drift in online health discussions
Term frequency-inverse Document Frequency (tf-idf)


• Article from Int’l Journal of Drug policy deploying tf-idf to identify Novel Psychoactive Substances (NPS) on reddit:
Word Embeddings

- Word2Vec paper
  https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

- State-of-the-art model, BERT
  - BERT model Python implementation https://huggingface.co/bert-base-uncased
  - BERT Paper https://aclanthology.org/N19-1423
  - BERT explanation http://jalammar.github.io/illustrated-bert

- Article from the British Journal on Social Psych about the application of word embeddings to capture cultural bias
Topic Modeling

• Readings for understanding topic modeling
  • Main paper https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

• Implementations
  • Java implementation (MALLET) https://mimno.github.io/Mallet/index
  • Python implementation (uses MALLET) https://radimrehurek.com/gensim_3.8.3/models/wrappers/ldamallet.html

• JMIR article exploring application of LDA and sentiment analysis to detect discussions on Twitter about the COVID-19 vaccine
Infodemiology as a Framework

• Journal of Medical Internet Research (JMIR) video & description of infodemiology:
  • https://infodemiology.jmir.org/

• Fully article by Gunther Eysenbach on infodemiology in the American Journal of Preventive Medicine:
Privacy as Contextual Integrity:

- Link to the original publication of Helen Nissenbaum’s 2004 Washington Law Review covering the topics of ”Privacy as Contextual Integrity”