Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Effect of Feedback on Drug Consumption Disclosures on Social Media

Hitkul Jangra[1], **Rajiv Shah**[1], **Ponnurangam Kumaraguru**[2]

[1]Indraprastha Institute of Information Technology, Delhi, India
[2]International Institute of Information Technology, Hyderabad, India
{hitkuli,rajivratn}@iiitd.ac.in, pk.guru@iiit.ac.in

Abstract

Deaths due to drug overdose in the US have doubled in the last decade. Drug-related content on social media has also exploded in the same time frame. The pseudo-anonymous nature of social media platforms enables users to discourse about taboo and sometimes illegal topics like drug consumption. User-generated content (UGC) about drugs on social media can be used as an online proxy to detect offline drug consumption. UGC also gets exposed to the praise and criticism of the community. *Law of effect* proposes that positive reinforcement on an experience can incentivize the users to engage in the experience repeatedly. Therefore, we hypothesize that positive community feedback on a user's online drug consumption disclosure will increase the probability of the user doing an online drug consumption disclosure post again. To this end, we collect data from 10 drug-related subreddits. First, we build a deep learning model to classify UGC as indicative of drug consumption offline or not, and analyze the extent of such activities. Further, we use matching-based causal inference techniques to unravel community feedback's effect on users' future drug consumption behavior. We discover that 84% of posts and 55% comments on drug-related subreddits indicate real-life drug consumption. Users who get positive feedback generate up to two times more drugs consumption content in the future. Finally, we conducted an anonymous user study on drug-related subreddits to compare members' opinions with our experimental findings and show that user tends to underestimate the effect community peers can have on their decision to interact with drugs.

Introduction

In 2019, 70,630 people died due to drug [1] overdose in the US alone; this number has almost doubled from 38,329 in 2010 (NIH 2021). The US president declared the drug crisis as a national public health emergency in 2017.[2]

A similar increase has also been observed in drug-related user-generated content on social media. The number of unique users in *r/Drugs* has gone up by 324% between 2012 and 2017 (Lu et al. 2019). Anonymity and limited content moderation make Reddit[3] an appealing platform for participating in unfiltered conversations on shared interests.

Though drug-related conversations on Reddit vary widely in their purpose, we are particularly interested in content that indicates offline drug consumption by a user.[4] These can be content where a user directly talks about their experience with consuming drugs, e.g., *Just downed this bad boy! 473mg tonight, wish me luck boys!* Sometimes content may not talk about a drug experience directly but indicate the intent of drug consumption, e.g., *I recently got two orange pyramid geltabs and was wondering if I should never handle them like tabs or if they are ok to touch a little bit.* These content pieces are interesting because they are online proxies for authors consuming drugs offline. Hereafter, we call user-generated content (post or comments) like these *drug consumption activity*.

An increasing amount of research has used Reddit to study various drug-related problems like drug abuse (Hu et al. 2019), forecasting drug overdose (Ertugrul, Lin, and Taskaya-Temizel 2020), transition into drug addiction (Lu et al. 2019), patterns of drug use and consumption methods (Balsamo et al. 2021), and geospatial patterns in drug use (Balsamo, Bajardi, and Panisson 2019). Though all these studies shine a light on the various patterns of drug consumption using digital data, none of them quantify the effect of the platform and community itself on drug consumption behavior. Research has shown online community feedback has an effect on multiple facets of users offline behavior like weight loss (Cunha, Weber, and Pappa 2017), physical activity (Althoff, Jindal, and Leskovec 2017), smoking and drinking relapses (Tamersoy, Chau, and De Choudhury 2017), quality of user-generated content (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014) and involvement in open-source projects (Valiev, Vasilescu, and Herbsleb 2018).

To fill this gap, we seek to quantify the effect of the platform and community on drug consumption behaviour. We collect data from 10 drug-related subreddit; develop a deep learning classifier to label activity as indicative of drug con-

[1]In this paper, the term "drug" represents illicit substances and not generic medical drugs.

[2]https://www.cms.gov/About-CMS/Agency-Information/Emergency/EPRO/Current-Emergencies/Ongoing-emergencies

[3]https://www.reddit.com

[4]Disclaimer: We do not oppose the existence or the way these subreddits function - as they can be helpful for support and harm reduction. Similarly, we do not view drug consumption negatively or condone it, as a sizable population might be indulging in it due to therapeutic or other social factors.

sumption or not, to quantify the extent of drug consumption activities. Further, grounded in *Primacy Effect* (Asch 1946) and *Operant Conditioning Theory* (Skinner 1938), we use propensity score matching (Stuart 2010) to quantify the impact of community feedback on the magnitude of future drug consumption activity posted by a user. Finally, we conducted an anonymous user study on our subreddits of interest to collect members' acknowledgment of drug consumption and opinion on the effect of community feedback on their subreddit and drug consumption behavior.

We discover that (1) deep learning classifiers can identify Reddit content indicative of drug consumption (macro F1 79.54), (2) 80.29% of users in drug-related subreddits have online activity indicating drug-consumption offline, which is in line with the response received in our user study, (3) 84.2% and 54.4% of all posts and comments posted on drug-related subreddits are indicative of drug consumption; (4) users' who receive positive feedback (comments or score) from the community on drug consumption activity tend to generate up to two times more drug consumption content in future, and finally (5) user's under-estimate the effect of community feedback can have on their decision to interact in drugs.

In summary, our main contributions are:

1. To reveal (using 10 subreddits) the causal effect online community feedback has on users' offline drug behavior.
2. A manually annotated dataset (4,000 samples) and deep learning classifier to detect UGC indicative of offline drug consumption.
3. An anonymous user study of drug-related subreddits members to compare community opinion with our statistical findings.

Our work impacts researchers, platform owners, and community moderators, providing a fertile base for developing harm-reduction research and tools. Our classifiers can be used to detect social media content indicative of drug consumption, providing opportunities for demographic-specific censoring or intervention. Our causal inference results and experiment setup can help platforms/communities design different methods of showing and providing feedback that can assist in harm reduction.

Data and Code: Reddit data is available via Pushshift API. [5] Our annotated dataset, user study responses, and modeling code is available at https://precog.iiit.ac.in/research/drug-feedback/.

Theories and Research Questions

Individuals prefer to present an idealized version of themselves; this phenomenon is known as *Impression Management* and is used to improve social standing among peers (Goffman 1959). Leary et al. (Leary, Tchividijian, and Kraxberger 1994) showed that individuals indulge in voluntary risk-taking activities like consumption of drugs, distracted driving, unprotected sex to improve impression among peers. Hogan (Hogan 2010) extends the concept of impression management to social media. He states that social media users can use status messages and media posted by them as a tool for impression management.

Subreddits are communities where having a positive impression/reputation can lead to various tangible and non tangible benefits like status, moderator privileges, Karma[6] and trophies. Thus we expect users could post drug consumption content to improve their impressions. Hence, we ask our first question:

RQ1. [Extent] *What is the extent (i.e. percentage of content, and users) of content indicating offline drug consumption in drugs-related subreddits?*

Our second research question is grounded in the *Primacy effect*, the tendency to remember the first piece of information (Asch 1946). For e.g., people's impression of an individual is dependent on the first traits they encounter (Asch 1946); probability of recalling initial items in a list is higher (Murdock Jr 1962); people have a more vivid memory of their first romantic encounter, achievements, and even losses (Dixit 2010). The primacy effect can cause *anchoring bias*, leading to skewed decisions relaying heavily on the initial information (Tversky and Kahneman 1974). Building on these theories, (Shteingart, Neiman, and Loewenstein 2013) proposed *outcome primacy*, proving long-lasting effects of the first experience. We hypothesize that the community feedback on the first drug consumption post can affect the user's future drug consumption and posting behavior.

RQ2. [First Experience] *How does the community feedback on first drug consumption post affect users' future drug consumption?*

Besides feedback on the first experience, user experience can also be dependent on *law of effect*, actions that are closely followed by satisfaction are more likely to reoccur (Thorndike 1898). Based on this principle, Skinner et al. proposed *Operant Conditioning* (Skinner 1938). It states the probability of acting in the future is a function of the outcomes received in the past. Positive reinforcement will incentivize the user to repeat an action in the future. Similar behavior is observed in the context of social media, e.g., more number of comments on post leads to higher weight loss (Cunha, Weber, and Pappa 2017), increased social media interactions lead to higher steps in activity tracking apps (Althoff, Jindal, and Leskovec 2017) and community feedback affects the quality of future posts (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014). Grounded in these theories, we expect continued positive feedback can affect a user's future drug consumption activity. We therefore ask:

RQ3. [Feedback] *How does continuous positive community feedback affect users' future drug consumption?*

Before studying the causal effect of feedback, we need to be able to detect drug consumption activity. An essential prerequisite to our work is building a classifier that can predict users' drug consumption in the offline world via user-generated textual content. A popular methodology in Natural Language Processing (NLP) is to learn dense rep-

[5] https://github.com/pushshift/api

[6] https://reddit.zendesk.com/hc/en-us/articles/204511829-What-is-karma-

resentations of text. Mikolov et al. (Mikolov et al. 2013) proposed a neural algorithm to learn text representations based on word co-occurrence, which outperformed classical token-based representation in a variety of classification tasks. Vaswani et al. (Vaswani et al. 2017) proposed an improved model architecture called Transformers based on self-attention (Shaw, Uszkoreit, and Vaswani 2018) to learn contextually aware dense text representations. Transformer-based large pre-trained models (Devlin et al. 2018; Liu et al. 2019) have provided an efficient base to perform classification on a variety of tasks and data sources. We build a deep learning classifier based on these architectures, asking:

RQ4. [Detection] *Can we use Reddit textual data to classify between drug consumption and non-drug consumption content? How accurate is such a classifier?*

Related Work

Our work is about the effect of social media community feedback on users' drug consumption behavior. Our related work flows from three directions - (1) Drug studies leveraging social media, (2) Causal inference using social media data, and (3) Self-harm behavior on social media.

Drug Studies on Social Media: Ease of data availability and many active communities around drugs have enabled a variety of related research. (Lu et al. 2019) built a machine learning classifier trained on textual features to identify users at risk of addiction and transition into drug recovery. They further use survival analysis to identify how much time it will take to undergo the transition. (Balsamo et al. 2021) used Word2Vec (Mikolov et al. 2013) similarity to curate a list of words used by Reddit users for different drugs, Routes of Administrations (ROA), and drug tampering techniques. Using the list, they rank the popularity of various drugs and ROAs. They report that between 2014 to 2018, the popularity of synthetic drugs like Fentanyl and unconventional ROAs like rectal administration of drugs has increased, whereas a decline has been observed in conventional ROAs like inhaling and injecting. (Balsamo, Bajardi, and Panisson 2019) filtered all activities of users on drug subreddits to extract location information and study the geospatial patterns of drug consumption in the US.

Besides Reddit, (Hu et al. 2019) used deep learning ensemble models to detect drug abuse in tweets and (Ertugrul, Lin, and Taskaya-Temizel 2020) used community attentive neural networks to forecast drug overdoes using information about crime dynamics.

A range of work has been done to understand the effect of peer groups on drug and alcohol consumption in offline work. (Farber, Khavari, and Douglass 1980; Kandel 1980) showed that people engage in alcohol, tobacco, and drug consumption since it can facilitate social image and peer acceptance. People indulge in substance consumption as it gives them an image of toughness, independence, maturity (Covington and Omelich 1988; Chassin et al. 1981; Camp, Klesges, and Relyea 1993) and makes them feel like they are "part of the gang" (Clayton 1991). However, no work has been done exploring the community's effect on drug consumption on social media platforms.

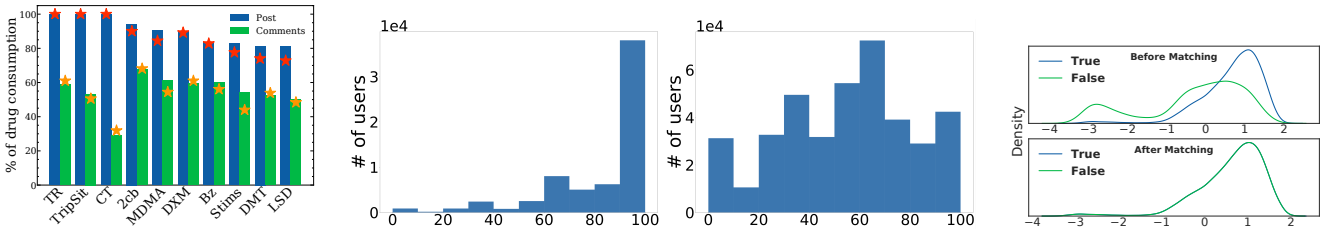
Causal inference using online data: Traditionally, researchers have established randomized controlled trials to establish causations. However, due to logistical and ethical concerns, such trials are not always feasible; e.g., it is not ethical and legal to make subjects consume illicit substances to study feedback's effect. For such studies, research has utilized publicly available online data. Additionally, the Internet provides a large volume of data, which is logistically impossible to obtain from controlled physical experiments. Careful filtration and analysis of large online data can help us simulate a randomized control trial (Rosenbaum and Rubin 1983).

(Cunha, Weber, and Pappa 2017) showed that positive feedback from the online community could help users lose more weight. (Althoff, Jindal, and Leskovec 2017) studied data from an exercise logging application and found increased social connections on the platform caused higher physical activity in the offline world. (Tamersoy, Chau, and De Choudhury 2017) used survival regression to establish a causal relation between linguistic cues from user-generated content and smoking or drinking relapse. (Kiciman, Counts, and Gasser 2018) used social media posting behavior to identify alcohol consumption and academic success of college students. Their analysis proved a causal relationship between high alcohol consumption and poor academic performance. (De Choudhury et al. 2016) unveiled the causal relation between user's vocabulary and suicidal tendency.

Social media data have also shown effects in opposition to expectations, e.g., (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014) showed that negative community feedback leads users to create even worst quality posts in the future rather than improving. Repercussions of feedback are topic and community dependent. Lack of literature analyzing feedback in drug and self-harm communities makes it an important area to study.

Self-harm behavior on social media: In the context of impression management, it has been shown that users tend to take part in self-harm activities like drug consumption and unsafe sex to improve social standings (Leary, Tchividijian, and Kraxberger 1994). An increasing number of users are getting involved in dangerous social media challenges like the KiKi Challenge (Baghel et al. 2018), the Salt and Ice Challenge (Roussel and Bell 2016), the Cinnamon Challenge (Grant-Alfieri, Schaechter, and Lipshultz 2013), Tide Pod Challenge (Murphy 2019), and the Fire Challenge (Ahern, Sauer, and Thacker 2015).

(Lamba et al. 2020) analyzed public Snapchat data from 173 cities around the world, revealing 23.5% of total 6.4 Million samples were examples of distracted driving. They performed demographic analysis to reveal that young males from the Middle Eastern and Indian subcontinent are more likely to produce distracted driving content. Similarly, (Nanda et al. 2018; Kurniawan, Habsari, and Nurhaeni 2013) analyzed deaths caused by taking selfies in dangerous situations like elevation, near water-bodies, or with firearms.



(a) Subreddits (b) % Posts (c) % Comments (d) Propensity Logits

Figure 1: (a) Perecentage of drug consumption content across subreddits. Values derived from proposed model are indicated by bars, and ⋆ shows values from manual annotation. (b)&(c) are distribution of % posts and % comments indicating real world drug consumption per user. (d) Distribution of propensity logits before (top) and after (bottom) matching.

Data Collection and Dataset

We use Reddit, a widely used social media platform. Reddit is formed by a collection of communities called *subreddits*. As of October, 2021, Reddit has 52 Million daily active users and 3 Million subreddits.[7] Subreddits are largely allowed to moderate their own community posts and the anonymity allowed, makes it a suitable platform for relatively unfiltered discourse compared to other social media platforms. Each subreddit is built around a specific topic. Users post content related to their interest and fellow users can *comment* on these posts, which creates a *thread*. Users can also *upvote* and *downvote* a post or comment, though only the total aggregate of votes is visible called *score*.

Reddit has several subreddits built around the topic of drugs. Wiki page of *r/Drugs* maintains a list of popular drug-related subreddits.[8] These subreddits contain different facades of drugs like addiction, recovery, cultivation, and experience. Some are drug agnostic like *r/tripreports* whereas others are drug specific like *r/MDMA* or *r/LSD*. We manually audited all the subreddits in the list and filtered 10 subreddits (see Table **??** in appendix), which is either (1) based around users sharing personal drug consumption experiences or (2) has a popular *flair*[9] indicating offline drug consumption.

To obtain the data from Reddit, we use the Pushshift API. For each subreddit, we collected all the threads made from the inception of the subreddit. Each thread contains the original post, the comments made, and scores for all activity in the thread. In total, we collected 826,905 posts and 6.6 Million comments made by 493,906 unique users. Only 269,059 unique users at least have one post; Table **??** in appendix provides a summary of statistics for each subreddit.

User Study Design

The impact of our research can be dependent on two factors: 1) Do the users actually consume drugs in real life, and 2) Analyzing causal inference results in light of members' perception since it can dictate the design of the effective intervention and education strategy.

To this end, we conduct a voluntary anonymous user study with members of 10 subreddits we are studying. Necessary permissions from the Institute's Review Board and moderators of subreddits were obtained before conducting the user study. Firstly, participants were asked to acknowledge (Yes or No) if they consumed drugs during their active period on the subreddit. Later, they were asked a series of questions about how much impact community feedback, number of comments, and score have on their future participation in subreddit and drug consumption. We wanted a quantitative understanding of users' perceptions rather than a simple yes/no answer while keeping the study's cognitive load low. Hence, we opted for the 5-point Likert scale (Likert 1932), 1 being *No Impact* and 5 being *Essential*. User study questioner can be accessed at https://forms.gle/yRqRriSPbgG9p2gN8. Total 45 users participated in our study. Results of each component are presented with the corresponding computational results.

Detecting Drug Consumption Content

To understand the extent of drug consumption behavior (**RQ1**), we first need to identify which user-generated content indicates drug consumption in real life (**RQ4**). Past research has assumed being active on drug-related subreddit as a proxy of drug consumption (Lu et al. 2019; Balsamo et al. 2021; Balsamo, Bajardi, and Panisson 2019). Though this may be true in most cases, users can also join the community as bystanders, for research purposes, or to help others. Hence, considering mere participation as a proxy of drug consumption is a weak assumption. Some subreddits have flairs that indicate drug consumption, but adding flairs to post is voluntary, and users may choose not to do so. Moreover, comments do not have a flair but still can indicate drug consumption. Towards solving this, we build a classifier that can mark posts and comments as indicative of drug consumption or not. Henceforth, we will use the term *activity* to represent user-generated posts or comments.

Ground Truth Annotation

To build a classification model, we need to have a ground truth dataset of activities labeled as drug consumption or non-drug consumption. The goal is to mark a sample as positive if it indicates the author consuming drug offline. We sample 4,000 user activities for annotation. To ensure a well-

[7] https://backlinko.com/reddit-users

[8] https://www.reddit.com/r/Drugs/wiki/subreddits

[9] https://www.reddit.com/r/help/comments/3tbuml/whats_a_flair/

distributed ground truth, half of the samples were posts, and half were taken from comments. Further, a uniform split is maintained across all 10 subreddits.

Annotation Guidelines: Annotators were provided with the text content and title (in case of posts) of an activity. An activity should be annotated as drug consumption in case of self-disclosure by the author, or if clear indication of author's possession/intent to consume drugs is present. For example:

- Self-disclosure: *Haha I had a bad trip off 30mg and weed first time but can't wait to try smaller doses.*
- Intent to consumption: *I'd be up for a distanced experience with a stranger (s) Just itching to get out of this awkward routine....*
- Drug possession: *I'm thinking about dissolving it in some alcohol and putting it in empty caps, not sure it will be better...*

It is important to note that just the presence of drug-related words does not imply drug consumption. The following examples contain drug-related words but do not indicate drug consumption:

- *My fourteen year old niece is smoking pot. What would /r/trees tell a fourteen year old about the effects of Marijuana? She might believe YOU and sources you cite.*
- *Mdma tolerance information Does anyone have any information on immediate mdma tolerance or articles about the subject*

Annotators were also given a list of drugs *street* names and slangs used in drug-related subreddits to assist the annotation process (Balsamo et al. 2021). Each sample was annotated by 3 annotators independently. We obtain a Fleiss-kappa (Fleiss 1971) agreement rate of 0.69, which signifies substantial agreement (Landis and Koch 1977). An activity was marked as drug consumption if 2 or more annotators agreed.

Dataset: 2,614 (65.32%) of 4,000 samples were marked as drug consumption, 79.35% of posts and 51.30% comments were marked as positive, respectively. Since comments are made in response to posts providing specific information, feedback, or expressing gratitude, a lesser positivity rate of drug indication than posts is expected. We make our annotated data public for future use. [10]

Deep Learning Classifier

We randomly split the manually labeled dataset into a train and test set of 3,200 (2,091 drug consumption, 1,109 non-drug consumption) and 800 (523 drug consumption, 277 non-drug consumption). Five-fold cross-validation is performed on train set to tune models, and final models are evaluated on the test set.

Model: Performing text classification with a combination of neural models and dense text representations has become a norm in NLP. Following the same, we experiment with different types of neural network models combined with contextual and non-contextual text embeddings. Our first model is a single channel one-dimensional convolutional neural network (Text-CNN) (Kim 2014). Input text for the Text-CNN model is vectorized using pre-trained Google News corpus Word2Vec embedding (Mikolov et al. 2013).

Transformer-based large pre-trained models with their ability to capture sentence context have achieved state-of-the-art performance on a variety of NLP tasks (Vaswani et al. 2017). Leveraging that, we experimented with BERT, a model built using bidirectional transformers and pre-trained on masked language model, and next sentence predictions tasks (Devlin et al. 2018). We also built a classifier based on RoBERTa (Liu et al. 2019), an optimized version of BERT. Table **??** reports 5-fold cross-validation performance of all the models.

Training Details: Our model is trained using Adam optimizer with the learning rate of $3X10^{-4}$, batch size 64, and utilized dropouts for regularization. We train models for 100 epochs with early stopping and checkpointing the best-performing model on the validation set. The training was performed on an Nvidia RTX 3090 GPU. Our code is available publicly for reproducibility and future use purposes.[10]

Validation and Robustness of Classifier: To further validate the generalizability of our models, we validate its performance on the test set (not used in the training step). Table **??** provides performance of all models on test set. Our best model achieve a macro F1 score of **79.54**. Table **??** in appendix provides performance numbers of our best model across subreddits.

Extent of Drug Consumption

We want to discover the extent of content on drug-related subreddits that indicates offline drug consumption by the user (**RQ1**). We use the proposed classifier to generate predictions for all the activities (posts or comments) that are not already marked as drug consumption by a flair or subreddit. We found that 84.2% of all posts and 54.4% of all comments indicate drug consumption by the user. Figure 1(a) shows the percentage of drug consumption posts and comments for each subreddit individually. $\star$ in the Figure indicates the drug consumption percentage observed in our manual annotation. A consistent slight difference between predicted and annotated drug consumption percentages shows the proposed model's robustness across subreddit and content types.

Once we have drug consumption prediction for all the activities, we aggregate them based on user ids and observe what % of users have activities for whom we have a positive prediction. We found that across 10 subreddits, 80.29% of all users in our dataset have consumed drugs. This is echoed in our user study findings too, where 84.4% participants (38 out of 45) acknowledged consumption of drug. As shown in Figure 1(b) $90\% - 100\%$ of posts for most users are indicative of drug consumption offline. The distribution of user's comments is less skewed, centered around the 60%-70% (Figure 1(c)). This proves a strong proxy between user activity on drug-related subreddit and drug consumption in the offline world and signifying the importance of studying the platform's impact on users' future online and offline activity.

[10] https://precog.iiit.ac.in/research/drug-feedback/

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
5-fold cross validation				
Text CNN	73.51 ± 3.10	78.79 ± 1.82	63.28 ± 5.78	62.34 ± 7.28
BERT	83.79 ± 1.03	82.13 ± 1.15	82.22 ± 1.29	82.14 ± 1.16
RoBERTa	83.16 ± 0.95	81.72 ± 1.32	80.96 ± 1.15	81.22 ± 0.98
Test set				
Text CNN	78.65	77.52	73.71	74.89
BERT	81.27	79.51	78.67	79.05
RoBERTa	**81.90**	**80.43**	**78.89**	**79.54**

Table 1: Drug consumption classification performance.

Observation 1 (Extent) *User-generated content of about 80% of total users in drug-related subreddits indicates drug consumption in real life. This is inline with the data received via our user study.*

Observation 2 (Extent) *84% user-generated posts and 54% comments indicate drug consumption. For majority user 90%-100% of their posts and 60%-70% of comments indicate offline drug consumption.*

Causal Analysis

In **RQ2** and **RQ3**, we aim to understand the causal effect that receiving positive feedback on drug consumption posts has on the users' future drug consumption activity. To this end, we use Propensity Score matching, a causal inference model shown to reduce bias compared to the naive correlation analysis (Imbens and Rubin 2015).

In the potential outcome framework (Neyman 1923), the "effect" of an experience on the outcome is formalized as an outcome $Y_i(T = 1)$ after a person i had the target experience T, i.e., treated,[11] and outcome $Y_i(T = 0)$ when the same person in the same circumstances has not received the treatment. The causal effect of the experience T is estimated as $Y_i(T = 1) - Y_i(T = 0)$. However, it is impossible to have the same individual receive and not receive treatment simultaneously. Propensity score matching attempts to overcome this challenge by observing the outcome on two different individuals, one treated and the other control but having similar treatment probability and confounders.

Feedback Threshold: In our case, treatment is the feedback received on a drug consumption post which is measured by the number of comments and scores [12] received. Conventionally, treatment is a binary variable (e.g., vaccine administered or not), and hence the assignment of treatment is trivial, but in our case, treatment is a continuous variable. We use hard thresholds (θ) to divide feedback into positive or negative and present results across various values of θ. It means, a post is considered to have positive feedback if it receives greater than θ number of comments or score. Averaged across our 10 subreddits, 80% of drug consumption activities receive less than 1.1 ± 0.3 comments and 2.9 ± 1.13 scores. To ensure robustness and generalizability in results, we experiment by varying our θ from 2 to 6, both inclusive.

Group Assignment: In **RQ2**, we analyze the treatment outcome on a user's first-ever drug consumption post. User is assigned to the treatment group if their first drug consumption post receives positive feedback. Additionally, in **RQ3**, we aim to study the effects of continuous feedback. A user at their n^{th} drug consumption post is assigned to the treatment group if all their past drug consumption posts, including n^{th}, have individually received positive feedback. We experiment with values of n between 1 to 6, both inclusive.

Propensity Model and Matching: After group assignment, we need to find pairs of users who have a similar likelihood of receiving treatment, but one is treated, and the other is not. In our case, given drug consumption post n and feedback threshold θ, propensity model estimates $P(n_{feedback} \geq \theta)$. Latent confounders encoded in linguistic and content characteristics, past feedback, and volumes can affect a post's feedback. In our experiment, we account for all these confounders while matching to create balanced treatment and control groups. Choices of our confounders are inspired by previous work using causal inference on social posts like (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014; Saha et al. 2019; Tamersoy, Chau, and De Choudhury 2017), and can be divided into 3 broad categories:

- *Content text*: User-generated textual content can give a measure of multiple confounders. In our case also, text of the drug consumption post is the main confounder and is being used for propensity score prediction.
- *Past activity*: Apart from text, we also use users' frequency of past activity as a confounder. While performing matching, only users with a similar number of posts and comments done in the past are paired together.
- *Past feedback*: Another important confounder regularly used in literature is the feedback (scores and comments in our case) received by a user in the past.

Multiple recent social media causal inference studies have used text-based models for propensity estimation (Cheng,

[11] In causal analysis literature, the subject who received the target experience is called treated and becomes part of the Treatment group. Whereas users who do not receive the target experience are referred as Control group.

[12] Score is an aggregate of number of up votes and down votes received. Only the aggregate is reported by Reddit not the individual values.

Danescu-Niculescu-Mizil, and Leskovec 2014; Sridhar and Getoor 2019; De Choudhury et al. 2016; Kiciman, Counts, and Gasser 2018; Saha et al. 2019). Most of these studies use a combination of n-gram features and Logistic Regression to train the propensity model (Keith, Jensen, and O'Connor 2020). However, recently (Weld et al. 2021) showed that choice of model architecture for text propensity model could induce bias in causal inference results. They experimented with a wide range of text representations (n-grams, LDA, contextual embeddings) and architectures (Logistic Regression, Simple NN, and BERT-derivatives) and found that BERT-based models were least prone to induce bias.

Considering (Weld et al. 2021) findings, we use pretrained RoBERTa (Liu et al. 2019) model for propensity estimations. Since subreddits may have different community dynamics and rules, separate models are trained for each subreddit across the range of feedback thresholds.[13] Size of training data was capped at 10,000 samples. An $80:20$ train test split was used for evaluation. Accuracy and macro F1 of our propensity models varied for subreddits between 57.8% to 89.2% and 43.3% to 70.1% respectively. It is important to note that a propensity model aims to build a descriptive selection model and not a predictive model (Rosenbaum and Rubin 1983), and hence, the importance of classification performance is secondary (Kiciman, Counts, and Gasser 2018). Further, (Weld et al. 2021) demonstrated that a highly accurate propensity model could induce bias in the estimation of the causal effect. Therefore, we move forward with propensity models having moderate performance.

Matching: A user in treatment group is matched with one user from control group when posts made by both have a similar propensity score. Generally, given a propensity score p, matching is done on $logit(p)$ (Equation 1). A pair is considered as a good match, if difference of $logit(p)$ is less than a *caliper* value as defined in Equation 2 (Harris and Horst 2016).

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{1}$$

$$caliper = 0.25 \times \sigma(logit(p)) \tag{2}$$

For a given treatment user, we filter all control users with $logit(p)$ difference less than $caliper$ value and then conduct a greedy search to find the nearest value. Matching is done in a one-to-many fashion.

Apart from propensity score, number of past activities and feedback received in past should also be balanced as confounders (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014). We ensure balance by matching the n^{th} drug consumption post made by both users, and treatment is only assigned if all the drug consumption posts from 1 to n individually receive positive feedback.

Quality of matching: Finally, to ensure the treatment and control group after matching are statistically similar, we use standardized mean difference (SMD) also known as Cohen's D (Stuart 2010) defined as: -

$$SMD = \frac{\overline{x}_{treatment} - \overline{x}_{control}}{\sqrt{\frac{\sigma^2_{treatment} + \sigma^2_{control}}{2}}} \tag{3}$$

Here, $\overline{x}$ and σ represent mean and standard deviation, respectively. To ensure matching quality SMD is preferred over p-value hypothesize testing since it conflates changes in balance with changes in statistical power (Stuart 2010).

In literature, where text propensity models are built on n-gram features, SMD balance check is conducted on n-gram vectors (Cunha, Weber, and Pappa 2017; Kiciman, Counts, and Gasser 2018; Saha et al. 2019). Since our propensity model is deep learning-based, we use feature vectors extracted from the last hidden layer of our model along with the frequency of user past activity to conduct a balance check (Kallus 2020; Louizos et al. 2017; Johansson, Shalit, and Sontag 2016). We evaluate the SMD distribution of feature vectors before and after matching for treatment and control users. A confounder is considered to be balanced if SMD is less than 0.25 (Stuart 2010).

Effect Size: Once we have our treatment and control groups statistically balanced upon confounders, effect of treatment can be calculated on the matched pairs. Estimated average treatment effect ($EATE$) is calculated as:-

$$EATE = \frac{\sum_{i=1,j=1}^{N} \frac{(Y_i(T=1) - Y_j(T=0)) * 100}{Y_j(T=0)}}{N} \tag{4}$$

$EATE$ gives an average percentage increase in the treatment group's outcome compared to the control group's outcome. Since the distribution of the treatment effect can be skewed, we report median values instead of mean.

Feedback on First Drug Consumption Post

We study the effect number of comments received by the first drug consumption post has on future drug consumption activity volume (**RQ2**). A user is assigned to a treatment or control group based on the number of comments received on their first drug consumption post. [14] We experiment with comment thresholds (θ) between 2 to 6.

We discover users who received positive feedback on first drug consumption post, generated upto 100% more drug consumption content in the future compared to the users in the control group. These results are statistically significant, evaluated using Kolmogorov-Smirnov test (Massey Jr 1951) and consistent across different treatment thresholds and subreddits. Table **??** shows $EATE$ of n_1 for all the subreddits calculated on $\theta = 4$. Figure 2 shows change in confounders SMD and Figure 1(d) changes in $logit(p)$ distributions before and after matching for ***r/LSD*** n_1, $\theta = 4$.

[13] Training parameters were similar to those presented in Section Detecting Drug Consumption Content. Training code is present in our code repository https://precog.iiit.ac.in/research/drug-feedback/

[14] Note that the first drug consumption post here represents the first post of the user which indicative of offline drug consumption in the subreddit. We do not claim this to be the user's first encounter with drugs in life.

	Comment ≥ 4						Score ≥ 4					
Subreddit	n_1	n_2	n_3	n_4	n_5	n_6	n_1	n_2	n_3	n_4	n_5	n_6
LSD	50.0^{***}	44.4^{***}	35.6^{***}	25.0^{***}	35.0^{*}	37.0^{**}	50.0^{***}	33.3^{***}	37.5^{***}	33.3^{*}	53.9^{*}	0.0
MDMA	75.0^{***}	52.9^{***}	50.0^{***}	27.6^{*}	50.0^{***}	71.4^{***}	41.4^{***}	53.8^{***}	50.0^{*}	41.1	64.0^{***}	129.9
Benzodi-azepines	75.0^{***}	50.0^{***}	35.0^{***}	52.7^{***}	33.3^{*}	30.0^{*}	50.0^{***}	75.0^{***}	66.7^{**}	133.3^{**}	183.3^{*}	266.6^{*}
Stims	82.6^{***}	63.6^{***}	42.8^{***}	40.0^{***}	37.5^{***}	68.4^{***}	38.4^{***}	30.0^{*}	51.9^{**}	12.3	-13.3	158.7^{**}
DMT	66.7^{***}	40.0^{***}	30.0^{***}	20.5^{*}	25.0^{***}	26.7^{*}	43.6^{***}	32.2^{*}	33.3^{*}	52.3	28.2	47.0
DXM	66.6^{***}	33.3^{***}	45.5^{***}	33.3^{***}	20.0	47.2	33.3^{***}	27.3	41.7	58.9	109.4^{*}	85.7
Currently tripping	60.0^{***}	100.0^{***}	255.0^{**}	31.6	465.3	1033.3	50.0^{***}	60.0^{***}	50.0^{***}	100.0^{***}	37.0	142.9^{*}
2cb	100.0^{***}	50.0^{***}	50.0^{*}	21.5	0.16	29.9	100.0^{***}	83.3^{***}	266.7	167.8	281.0	206.4
TripSit	80.0^{***}	50.0^{**}	33.5	14.3	25.0	17.5	33.3^{*}	50.0	-9.4	276.3	20.0	N/A
TripReports	100.0^{***}	44.4	41.7	21.4	-62.5	N/A	14.3	150.0	350.0	233.3	N/A	N/A

Note: $^{***}p <= .001$, $^{**}p <= .01$, $^{*}p <= .05$. N/A means no matching pairs for the configuration.

Table 2: $EATE$ of feedback threshold (θ) 4 on the number of future drug consumption activities. n_i represents the i^{th} drug consumption activity done by an user. Positive feedback consistently leads to a higher volume of future drug consumption activity. Lack of enough treatment users lead to statically insignificant results in some configurations.

Continuous Feedback on Drug Consumption Posts

Additionally, we check the causal effect when a user continuously receives positive feedback on drug consumption posts (**RQ3**). We repeat the matching experiments to evaluate the $EATE$ of the same outcome when the user receives consecutive positive feedback on their first n drug consumption posts i.e. all 1 to n drug consumption posts got positive feedback individually. Averaged across our 10 subreddits, we observe 80% of the users posts less than 6.9 ± 2.5 drug consumption activities in our time of observation. We experiment with values of n between 2 to 6. Table **??** shows the results for $\theta = 4$. We observe that treated users performed a higher number of drug consumption activities in the future. Our results are statistically significant. However, we do get insignificant results for experiment configurations with high values of θ and n due to the lack of enough matching pairs. This is more pronounced in smaller subreddits. However, we never receive a statistically significant result that conflicts with our hypothesis.

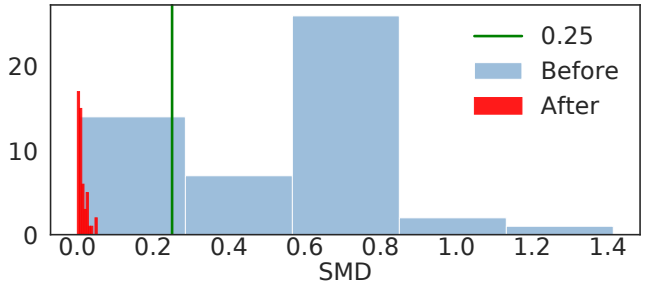


Figure 2: Matching quality for r/LSD, n_1, $\theta = 4$. Distribution of confounders' SMD before and after matching. After matching SMD for all confounders in ≤ 0.25 indicating good quality matching.

Score as Feedback

We also conduct all configurations of our experiments with the score as the treatment variable. Just as with comments, we receive consistent and statistically significant results; an increase in future drug consumption activity for treated users. Table **??** show results for $\theta = 4$.

To ensure robustness we experiment across a wide range of parameters ($\theta = [2, 6], n = [1, 6]$, comments and score as feedback) for each subreddit, leading to $\approx$ 600 experiment configurations. Due to lack of space, it is not feasible to present results of all the configurations in the paper. Results across the various experiment configuration are inline with our hypothesis and statistically significant. Complete results and statics of matching quality (before and after confounder SMD distributions), $EATE$, number of treatment control pairs, and statistical significance across all configurations are available at https://precog.iiit.ac.in/research/drug-feedback/.

Observation 3 (Increased Volume) *Positive community feedback on drug consumption posts (first and continuous) causes an increase in future drug consumption activity.*

Though causal inference shows a significant impact of community feedback on users' future participation and drug consumption, the impression of community members in our user study differs. Participants, on average, reported a *little* to *moderate* impact of community feedback on their behavior. On a 5 point Likert scale (1=*No Impact*, 5=*Essential* the average response was $2.28/5$ for scores and $2.53/5$ for comments. Such phenomenon of users under-estimating the effect of external factors on their participation in self-harm activity to maintain an "illusion of control" is well studied in the social psychological theory Layng's edgework (Lyng 1990). Understanding the contrast between user opinion and statistical findings is vital to designing effective intervention and harm-reduction strategies.

Observation 4 (Effect Underestimation) *Users on drug-related subreddits tend to underestimate the effect commu-*

nity feedback has on their future engagement and drug consumption.

Discussion

Research Questions

We begin our analysis with **RQ1** which aims to understand the extent of content in drug-related subreddits indicating drug consumption by a user in the offline world. Such content pieces provide a strong proxy for online-offline interaction of drug consumption and help quantify the prevalence of such self-harm behavior on social media. We discover that 84.2% of all posts and 54.4% of all comments posted on our observed subreddits indicate offline drug consumption. According to our model predictions, 80% of users have indulged in drug consumption, which is in line with the user acknowledgment we obtained from our user study. This distribution is consistent across subreddits irrespective of the subreddits theme (drug experience or not) or drug type. In fact, for most users, between 80% to 100% of their posts indicate drug consumption.

Primacy effect is a cognitive bias that explains people's tendency to depend on first experiences and impressions while making decisions. We validate does primacy effect holds for the users of drug-related subreddits. For social media users, feedback from the community can provide tangible and intangible benefits like gratification, a sense of belonging, special moderator status in the community. Thus in **RQ2**, we use propensity score matching to infer the causal effect positive feedback on first drug consumption post has on future drug consumption. Validated across different thresholds, we found that users who receive a high number of comments on first drug consumption post showed up to 100% increase in drug consumption indicative activity in the future.

Operant conditioning framework further expands the effect of feedback stating positive reinforcements can lead to repeated actions and habit building. In **RQ3**, we validate this by expanding our causal inference experiments to include continuous feedback received on drug consumption posts generated later in the timeline. We observe, similar to the first experience, receiving a continuous positive community feedback on drug consumption posts leads to an increase in magnitude of drug consumption activity. Observing **RQ2** and **RQ3** in tandem, we hypothesize that the feedback on the first drug consumption post can act as a "gateway" for the user; continuous feedback on later instances "reinforces" the habit. Together, positive feedback incentivizes a user to produce higher volumes of drug consumption content and, as a proxy, increased self-harm in the offline world.

Our user study unveiled, users perception of community feedback's impact on their behavior is less than what is observed statistically. Discrepancies like this have been studied in psychology literature (Lyng 1990) and can pose a danger to users well-being.

Finally, to answer our research questions, we need to classify subreddit activities (posts or comments) into indicative of drug consumption or not. Leveraging the large scale data available, to answer **RQ4** we train a deep learning classifier capable of classifying activities into offline drug consumption or not with high precision and recall. We further validate the robustness of the proposed model by evaluating performance on the test set spread across subreddits.

Implications and Ethical Considerations

All subreddits involved in our work list harm reduction as one of the community's primary goals. We believe our models and findings have direct implications for community moderators and platform designers involved in harm reduction interventions.

Feedback based: One of our key insights is increased drug consumption activity by users who received positive community feedback. Thus communities can experiment with different strategies of showing feedback, like only showing counts, partial, or rate limited feedback and quantify the reduction in said effect. Our insight and models can also help design community feedback guidelines regarding limiting community interactions on specific activities.

Intervention based: User's feedback history combined with our proposed deep learning classifier can help in monitoring drug consumption activity at an user or cohort level. High-risk individual(s) can be detected, and timely interventions like notifying, community reach outs, or restricted activity can help in reducing overall self-harm.

Some interventions may also have adverse effects; hence, more experimentation is required before moving forward. We acknowledge that tracking user data and restricting platform usage patterns can violate privacy and freedom of expression. However, our work does not aim at providing specific intervention methods. Instead, we provide necessary insights, data, and models that researchers and community moderators can use for further work based on every community's rules and ethics.

Resource based: A variety of research can be conducted on these platforms to understand and prevent the harms caused by drug consumption. However, the validity of any such work is dependent on ensuring that the online content provides a strong proxy for offline drug consumption. We open-source a manually annotated dataset and our pretrained models from drug consumption classification to enable further research.

Threats to Validity

It is always challenging to ensure generalizability while analyzing pseudo anonymous online data. Our analysis is also susceptible to these challenges. Firstly, our data is collected through Reddit, which can have biased representations in terms of geography, gender, and age. Further, though we experiment with 10 different drug-related subreddits varying across size, time, drug, and community objective, some other subreddits or social media platforms may not follow our insights. Finally, the users posting about drug consumption online may themselves not be a fair representation of the population engaging in drug consumption. However, since these people are consuming drugs and publicly generating content about it, we believe it is an important demographic

to study if we aim to understand the online-offline connection of drug consumption behavior.

In our analysis, user-generated drug consumption content is used as a proxy for offline drug consumption by the user. Since our data source is online, we do not have any way to ensure that the user did consume the drugs. We use data spread across various communities and long timelines adding up to millions of activities reducing the possibility of large scale tampered data. Further we perform a voluntary and anonymous user study in same communities to get acknowledgment of drug consumption. Our analysis and user study responses are based on the belief that users are not putting out false experiences. Additionally, it is necessary to note that the absence of online drug consumption content is insufficient for proving users not congesting drugs offline. Our study does not aim to make conclusions about drug consumers who do not actively interact with the platform.

Our experiments do not account for the sentiment of comments received to prevent errors in sentiment identification propagating to causal inference results. Due to drug/self-harm content dynamics, off-the-shelves sentiment models can cause unforeseen biases. A potential future work can be to train topic-specific sentiment models and observe their effect on the outcome.

Additionally, we control multiple contents, user, and community confounders while setting up our causal inference pipeline. However, there is always a possibility of unaccounted variables leaking into the causal inference outcomes. Finally, the sample size of our user study is small. Though this does not affect the primary statistical findings of our work, a more extensive and exhaustive study is desirable.

Conclusion

Our study investigates user-generated content indicative of drug consumption in the offline world. Specifically, we collect publicly available data from 10 drug-related subreddits and analyze the extent of drug consumption activity in these communities. First, we build a text-based deep learning model to classify user activities into drug consumption or not. Adapting from the sociology literature of feedback, we aim to test if the theories proposed for the offline world are also applicable to the behavior of posting drug consumption content on a social media platform. We put forth multiple RQs related to feedback's extent and causal effect on such behavior.

In summary, we observe that the majority of content posted on drug-related subreddits indicates drug consumption in the offline world. Further, we discover that users who receive positive community feedback on drug consumption content tend to generate higher volumes of similar content in the future, though users seem to underestimate this effect as shown by our user study. We believe that the observation made in our work can help to design online feedback mechanisms and interventions to reduce self-harm.

Appendix

Subreddit	# of Post	# of Comments	# of Users	# of Users with Post
LSD	343,346	2,658,323	266,185	138,073
MDMA	113,030	1,022,810	103,900	55,149
Bz	107,264	794,141	55,823	32,887
Stims	84,049	710,692	51,848	24,440
DMT	81,860	753,570	79,215	35,005
DXM	60,555	486,052	30,989	18,795
CT	17,388	50,757	19,540	6,650
2cb	9,258	83,642	11,348	5,317
TripSit	7,780	76,329	16,267	5,609
TripReports	2,148	10,991	3,791	1,659

Table 3: Statistics about the data collected.

Subreddit	Accuracy	Macro Precision	Macro Recall	Macro F1
TripSit	88.24	88.77	88.24	88.19
DMT	89.53	88.14	87.35	87.73
Stims	84.72	84.70	83.28	83.82
CT	82.35	81.60	84.47	81.79
LSD	82.93	82.11	81.21	81.60
2cb	84.71	78.58	81.35	79.77
DXM	85.86	77.92	81.80	79.55
TripReports	78.57	78.15	78.47	78.26
Bz	82.35	86.24	71.17	74.09
MDMA	76.92	74.44	69.30	70.68

Table 4: Performance of proposed model across subreddits. Sorted by Macro F1 score.

Acknowledgments

Hitkul is is supported by TCS Research Scholar Program. We also thank Hemank Lamba, and members of Precog Research Lab at IIIT-Hyderabad for their help and insights during this work.

References

Ahern, N. R.; Sauer, P.; and Thacker, P. 2015. Risky behaviors and social networking sites: how is YouTube influencing our youth? *Journal of psychosocial nursing and mental health services*, 53(10): 25–29.

Althoff, T.; Jindal, P.; and Leskovec, J. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *WSDM*.

Asch, S. E. 1946. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3): 258–290.

Baghel, N.; Kumar, Y.; Nanda, P.; Shah, R. R.; Mahata, D.; and Zimmermann, R. 2018. Kiki kills: Identifying dangerous challenge videos from social media. *arXiv preprint arXiv:1812.00399*.

Balsamo, D.; Bajardi, P.; and Panisson, A. 2019. Firsthand opiates abuse on social media: monitoring geospatial patterns of interest through a digital cohort. In *The World Wide Web Conference*.

Balsamo, D.; Bajardi, P.; Salomone, A.; and Schifanella, R. 2021. Patterns of Routes of Administration and Drug Tampering for Nonmedical Opioid Consumption: Data Mining and Content Analysis of Reddit Discussions. *Journal of Medical Internet Research*.

Camp, D. E.; Klesges, R. C.; and Relyea, G. 1993. The relationship between body weight concerns and adolescent smoking. *Health psychology*, 12(1): 24.

Chassin, L.; Presson, C. C.; Sherman, S. J.; Corty, E.; and Olshavsky, R. W. 1981. Self-images and cigarette smoking in adolescence. *Personality and Social Psychology Bulletin*, 7(4): 670–676.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2014. How community feedback shapes user behavior. In *ICWSM*.

Clayton, S. 1991. Gender differences in psychosocial determinants of adolescent smoking. *Journal of School Health*, 61(3): 115–120.

Covington, M. V.; and Omelich, C. L. 1988. I Can Resist Anything But Temptation: Adolescent Expectations for Smoking Cigarettes 1. *Journal of Applied Social Psychology*, 18(3): 203–227.

Cunha, T.; Weber, I.; and Pappa, G. 2017. A warm welcome matters! the link between social feedback and weight loss in/r/loseit. In *Proceedings of the 26th International Conference on World Wide Web Companion*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the CHI 2016*, 2098–2110.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dixit, J. 2010. Heartbreak and home runs: The power of first experiences. *Psychology Today*, 43: 60.

Ertugrul, A. M.; Lin, Y.-R.; and Taskaya-Temizel, T. 2020. CASTNet: community-attentive spatio-temporal networks for opioid overdose forecasting. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*, 432–448. Springer.

Farber, P. D.; Khavari, K. A.; and Douglass, F. M. 1980. A factor analytic study of reasons for drinking: Empirical validation of positive and negative reinforcement dimensions. *Journal of Consulting and Clinical Psychology*, 48(6): 780.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Goffman, E. 1959. The Presentation of Self in. Butler, Bodies that Matter. *The Presentation of Self in. Butler, Bodies that Matter*.

Grant-Alfieri, A.; Schaechter, J.; and Lipshultz, S. E. 2013. Ingesting and aspirating dry cinnamon by children and adolescents: the "cinnamon challenge". *Pediatrics*, 131(5): 833–835.

Harris, H.; and Horst, S. J. 2016. A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research, and Evaluation*, 21(1): 4.

Hogan, B. 2010. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*.

Hu, H.; Phan, N.; Geller, J.; Iezzi, S.; Vo, H. T.; Dou, D.; and Chun, S. A. 2019. An Ensemble Deep Learning Model for Drug Abuse Detection in Sparse Twitter-Sphere. In *MedInfo*.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Kallus, N. 2020. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, 5067–5077. PMLR.

Kandel, D. B. 1980. Drug and drinking behavior among youth. *Annual review of sociology*, 235–285.

Keith, K.; Jensen, D.; and O'Connor, B. 2020. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5332–5344. Online: Association for Computational Linguistics.

Kiciman, E.; Counts, S.; and Gasser, M. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 EMNLP*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

Kurniawan, Y.; Habsari, S. K.; and Nurhaeni, I. D. A. 2013. Selfie culture: Investigating the patterns and various expressions of dangerous selfies and the possibility of government's intervention. *The 2nd Journal of Government and Politics*, 324.

Lamba, H.; Srikanth, S.; Pailla, D. R.; Singh, S.; Juneja, K. S.; and Kumaraguru, P. 2020. Driving the last mile: Characterizing and understanding distracted driving posts on social networks. In *Proceedings of the ICWSM*, volume 14, 393–404.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Leary, M. R.; Tchividijian, L. R.; and Kraxberger, B. E. 1994. Self-presentation can be hazardous to your health: Impression management and health risk. *Health Psychology*.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Lu, J.; Sridhar, S.; Pandey, R.; Hasan, M. A.; and Mohler, G. 2019. Investigate transitions into drug addiction through text mining of Reddit data. In *KDD*.

Lyng, S. 1990. Edgework: A Social Psychological Analysis of Voluntary Risk Taking. *American Journal of Sociology*, 95(4): 851–886.

Massey Jr, F. J. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253): 68–78.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Murdock Jr, B. B. 1962. The serial position effect of free recall. *Journal of experimental psychology*, 64(5): 482.

Murphy, R. H. 2019. The rationality of literal tide pod consumption. *Journal of Bioeconomics*, 21(2): 111–122.

Nanda, V.; Lamba, H.; Agarwal, D.; Arora, M.; Sachdeva, N.; and Kumaraguru, P. 2018. Stop the killfies! using deep learning models to identify dangerous selfies. In *Companion Proceedings of the The Web Conference 2018*, 1341–1345.

Neyman, J. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10: 1–51.

NIH. 2021. Overdose Death Rates. *National Institutes of Health*.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Roussel, L. O.; and Bell, D. E. 2016. Tweens feel the burn:"salt and ice challenge" burns. *International journal of adolescent medicine and health*, 28(2): 217–219.

Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kıcıman, E.; and De Choudhury, M. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 440–451.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Shteingart, H.; Neiman, T.; and Loewenstein, Y. 2013. The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2): 476.

Skinner, B. F. 1938. The behavior of organisms: an experimental analysis. In *Appleton-Century*.

Sridhar, D.; and Getoor, L. 2019. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Tamersoy, A.; Chau, D. H.; and De Choudhury, M. 2017. Analysis of smoking and drinking relapse in an online community. In *Proceedings of the 2017 international conference on digital health*, 33–42.

Thorndike, E. L. 1898. Animal intelligence. *Nature*, 58(1504): 390–390.

Tversky, A.; and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157): 1124–1131.

Valiev, M.; Vasilescu, B.; and Herbsleb, J. 2018. Ecosystem-level determinants of sustained activity in open-source projects: A case study of the PyPI ecosystem. In *ESEC/FSE*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Weld, G.; West, P.; Glenski, M.; Arbour, D.; Rossi, R.; and Althoff, T. 2021. Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference. arXiv:2009.09961.