INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Modeling Online User Interactions and their Offline effects on Socio-Technical Platforms

Comprehensive Report submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

By

Hitkul

Under the supervision of

Dr. Rajiv Ratn Shah, IIIT Delhi
Prof. Ponnurangam Kumaraguru, IIIT Hyderabad

Indraprastha Institute of Information Technology Delhi

January, 2023

Abstract

Do online interactions trigger reactions back in the offline world? How can these reactions be detected and quantified? Specifically, what insights can be extracted for users, platform owners, and policymakers to minimize the potential harm of such reactions?

Society functions based on the complex interactions between individuals, communities, and organizations. The advent of the Internet has enabled these interactions to move online. A website or an application that facilitates the digitization of social interactions is called a *socio-technical* platform. For instance, individuals converse with each other via direct messaging applications (e.g., WhatsApp, Telegram), share thoughts, and gather feedback from communities (e.g., Reddit, Twitter, Youtube). Trade of goods occurs via e-commerce (e.g., Flipkart, Amazon) and online marketplaces (e.g., Google Play store). At times interactions happening in the online world, trigger reactions in the offline world, which we call *overflow*. Such overflows can have either a positive or negative impact. Socio-technical platforms save every interaction and associated metadata, providing a unique opportunity to analyze rich data at scale. Discover interaction patterns, detect and quantify overflow of interactions, and extract insights for users and policymakers.

This report aims to study the interactions by keeping the individual as the focal point. We focus on two broad forms of interactions - i) the effect online community feedback can have on individual offline actions and ii) organizations leveraging individual customers' online presence to optimize business processes. In the first part, we work on two scenarios - (a) How does community feedback affect an individual future drug consumption frequency in a drug community forum? and (b) What changes does an individual undergo immediately after getting sudden popularity in Online social media? What actions help in maintaining popularity for longer? In the second part, we leverage online information about a customer to improve the prediction of Return-to-Origin [1] in the e-commerce platform.

[1] https://easyinsights.ai/blog/return-to-origin-rto-why-is-it-a-crucial-metric-for-ecommerce-businesses/

Contents

Chapter 1

Introduction

Humans are inherently social. Life and society are structured around complex interactions. We communicate with each other to build family, friendship, and romantic relationships; to seek or provide advice and education; to execute trade and commerce. People unite to form organizations that drive economic activity, govern states, and provide social benefits. Human interactions are multidimensional, and different methods are used to structure these interactions based on the study area. As part of our work, we categorize interactions based on the entities involved. Our three interest entities are 1) Individuals, 2) Communities, and 3) Organizations.

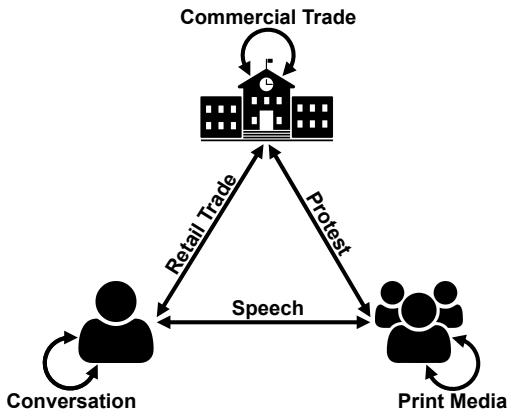


Figure 1.1: Different combination of societal interactions.

Figure 1.1 demonstrates different combinations of interactions that can take place between such entities. Individuals indulge with other individuals in private conversations, trade with organizations, and impart their thoughts to a community via speeches. Communities display their objection towards organizations via protests and help other communities as part of social service. Similarly, organizations interact with individuals, communities, and other organizations by conducting business. Researchers in the fields of Sociology, Psychology, Cognitive, and Economic sciences have long been interested in understanding the dynamics and characteristics of these interactions. The resulting literature has helped us design better communication structures, effective and efficient public policy, and derive economic growth.

1.1 Advent of Digital Interactions

An increase in the accessibility of the Internet has enabled a variety of *Socio-Technical* platforms. In our work, we define socio-technical platforms as a website or an application that facilitates the digitization of social interactions, e.g., WhatsApp, Twitter, Flipkart, and BidAssist. Focus of the report is to study the interaction facilitated by socio-techincal plat-

forms. Hence we start by describing some categories and examples of the same.[1] Figure 1.2 shows offline interactions and their corresponding socio-technical platforms.

- **Individual - Individual**: Platforms that allow users to have private conversations with others, like WhatsApp, Telegram, and Facebook Messenger.
- **Individual - Community**: Platforms that allow users to share their thoughts to a specific community (e.g. Reddit, Team-BHP, Facebook), or to a broad audience (e.g. Twitter, Moj, Youtube). Typically these platforms would also have feedback mechanisms (like, share, comments) for the community to react.
- **Individual - Organization**: Platforms that facilitate trade between individuals and an organization. For example, e-commerce platforms like Amazon and Flipkart; Online education platforms like Byjus and Coursera; and servicing platforms like Uber, Swiggy, and Dunzo.
- **Community - Organization**: A platform designed to facilitate communication between a large cohort of people with specific organizations like Change.org.
- **Organization - Organization**: Platforms are designed for organizations to interact with each other mainly to facilitate trade. For example, B2B e-commerce like BigCommerce, and Moglix; and Tender search engines like BidAssist.

Such platforms have transformed the range, scale, and efficiency of interactions. These platforms affect how we shop, vote, protest, and conduct our social life. All such platforms save every interaction on their system, including the entities involved, time stamps, and related follow-up interactions. The global scale of these platforms provides us with vast and rich data sources that were rarely available before their existence. The availability of

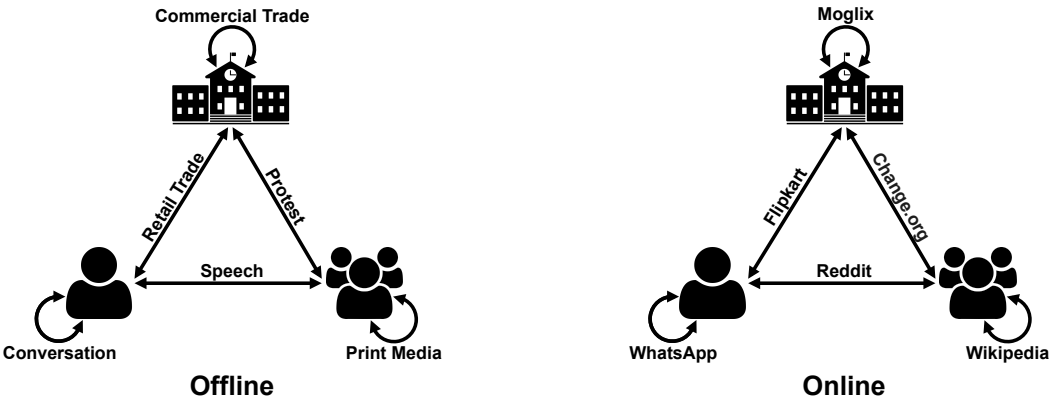


Figure 1.2: Offline interactions and corresponding online platforms.

[1]As platforms grow, it can be challenging to classify them into a single category. For Example, Twitter was designed as a public micro-blogging site but now also supports a direct messaging feature. Categorization in this report is based on the primary function of the platform.

such rich data provides unique opportunities to discover user patterns, validate old sociology theories at scale, and observe the change in behavior due to the omnipresent socio-technical platforms. Businesses can use this data to find and target customers. Insights from this data can be crucial for platform owners and governments to design engaging systems and effective public policies.

1.2 Online-Offline Interactions

We discussed various types of offline interactions and their online counterparts. However, the two worlds are not mutually exclusive. Feedback and trends from online interactions regularly trigger reactions in the offline world, where individuals and organizations morph their actions to achieve a favorable online persona. In our work, we call the phenomenon of online reactions affecting the offline world as *overflow*. Sometimes, such overflows can lead to positive changes like alternative career options (e.g., content creator), monitory growth, and increased reach/awareness. On the other hand, overflows can lead to profound negative implications like self-harm (KiKi challenge, Blue Whale challenge), financial fraud, and social unrest. To maximize the positive overflow and minimize the negatives, it is important to study i) the loop between online interactions and offline actions, ii) devise algorithms to detect and quantify the overflow, and iii) suggest measures for involved entities, platforms, and policymakers.

In our work, we aim to study the interactions by keeping the individual as the focal point. The report, focuses on two broad forms of interactions - i) the effect online community feedback can have on individual offline actions, and ii) organizations leveraging individual customers' online presence to optimize business processes. In the first part, we work on two scenarios - (a) How does community feedback affect an individual future drug consumption frequency in a drug community forum? and (b) What changes does an individual undergo immediately after getting sudden popularity in Online social media? What actions help in maintaining popularity for longer? In the second part, we leverage online information about a customer to improve the prediction of Return-to-Origin in the e-commerce platform.

1.3 Contributions

Individual - Community

- Gurjar, O., Bansal, T., Hitkul, Lamba, H., and Kumaraguru, P. Effect of Popularity Shocks on User Behavior. *In Proceedings of the 16th AAAI International Conference on Web and Social Media (ICWSM' 22), June 6-9, 2022, Atlanta, Georgia, USA*.
- Hitkul, Shah, RR., and Kumaraguru, P. Effect of Feedback on Drug Consumption Disclosures on Social Media. *R&R at top tier conference*.

Individual - Organization

- Hitkul, Abinaya, Saha, S., Banerjee, S., Chelliah, M., and Kumaraguru, P. Social Re-Identification Assisted RTO Detection for E-Commerce. *Under Review at a top tier conference*.

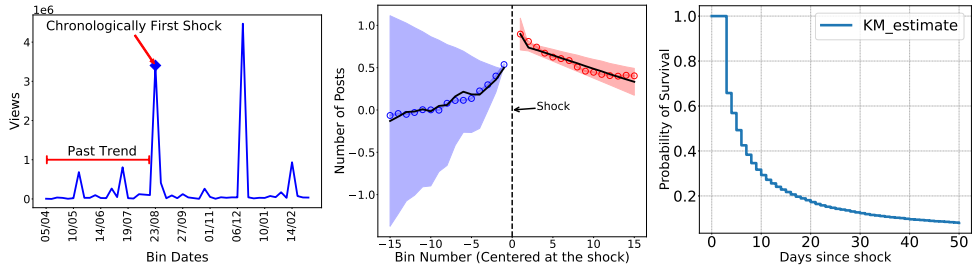
Chapter 2

Effect of Popularity Shocks on User Behavior

Users often post on content-sharing platforms in the hope of attracting high engagement from viewers. Some posts receive unusual attention and go “viral", eliciting a significant response (likes, views, shares) to the creator in the form of *popularity shocks*. Past theories have suggested a sense of reputation as one of the key drivers of online activity and the tendency of users to repeat fruitful behaviors. Based on these, we theorize popularity shocks to be linked with changes in the behavior of users. In this paper, we propose a framework to study the changes in user activity in terms of frequency of posting and content posted around popularity shocks. Further, given the sudden nature of their occurrence, we look into the survival durations of effects associated with these shocks. We observe that popularity shocks lead to an increase in the posting frequency of users, and users alter their content to match with the one which resulted in the shock. Also, it is found that shocks are tough to maintain, with effects fading within a few days for most users. High response from viewers and diversification of content posted is found to be linked with longer survival durations of the shock effects. We believe our work fills the gap related to observing users' online behavior exposed to sudden popularity and has widespread implications for platforms, users, and brands involved in marketing on such platforms.

2.1 Introduction

Recently, social media platforms have emerged or transformed themselves to focus more on content creation and sharing, e.g., TikTok, Instagram, Twitch, YouTube, etc. These social media platforms, focusing on content/multimedia sharing, have enabled users to express themselves in unique ways (text, photos, videos, etc.) to their followers (subscribers). To continue content creation and also engagement, most of the platforms have also launched creators' funds and also allow content creators (users) to get incentives/money to create such content [58, 125]. With social media content creation becoming an alternate source for revenue generation, users are also focusing on creating exciting content and eliciting attention and thus engagement from other users.



(a) Example of Shock (b) RDiT (c) Survival Analysis

Figure 2.1: Our work discusses (a) detecting points of sudden increase in response known as *popularity shocks* on users' timelines; (b) Quantifying behaviour change due to popularity shock in terms of change in posting frequency using RDiT(Regression Discontinuity in Time); (c) Short-lived survival duration of effect of shocks and factors affecting it.

Users who become popular on these social media platforms are often termed as "influencers" or "micro-celebrities" [125, 32, 49]. Influencers, due to their popularity, have a broad reach and have been studied in the past on swaying/forming attitudes about consumer purchase intention [69], brand's image [42] and perceived uniqueness [24], along with even dietary behavior in children [113]. Influencers are also often contacted by different brands for endorsing their products [125, 122]. Many studies have been done in understanding why certain content and users who post them become popular [29, 33, 30, 76]. However, there has been little or no study on addressing how viral users respond to their newly achieved popularity.

On content sharing platforms, receiving sudden popularity due to specific content (or a series of content getting viral) can be termed as *popularity shocks*. Popularity shocks can be characterized as a sudden increase in feedback (i.e., views, likes, etc.). Previously, popularity shocks have been studied towards Wikipedia pages because of an associated event [132, 53], and Github repositories due to being highlighted by the platform [73]. However, the effect of popularity shocks on users' content creating behavior has not been studied in detail. Similarly, much work has been done in predicting posts that will go viral or will become popular using initial dynamics [28, 133, 129]. However, little work has been done in analyzing the after-effects of a post becoming viral or a user becoming popular. *Do users become more active on the platform after getting popular? Do users alter their content or stick to the content that made them popular? How long does the popularity shock last? Is popularity short-lived, or can it be long-term based on how user conducts themselves?* Answering these questions could have wide-reaching implications for all three - the users, potential brands seeking influencers to partner with, and also the social media platform itself. Studying users' response to popularity shock can be insightful for (a) users, who want to continue engagement, (b) brands, for identifying new influencers which align with their values, and (c) social media platforms, for guiding new popular users on specific interventions that can be related to education, design changes or guidelines.

We ground our work in sociological theories related to social reinforcement and a sense of reputation. A reputable theory in the field of behavioral psychology has been *Operant Conditioning* [111]. Under this theory, an activity that earns rewards prompts an individual to repeat that activity, and similarly, an activity that earns punishment makes the individual more inclined to repeat that activity. In our context, if we treat receiving popularity, which is quantified with high engagement from the community on users' content, as positive feedback (or reward), the user ideally will keep repeating the same behavior. Alternatively, if the user received a popularity shock in a negative context, i.e., they were a recipient of a firestorm [60], they might stop posting similar content. We also draw on the theoretical work carried out in a more specific context of online communities [57, 98]. In one of the earliest analyses of an online community, Rheingold hypothesized that desire for prestige is one of the key motivators for individuals' contribution to the community. Kollack re-emphasized this [57], highlighting that increased reputation is one of the three reasons for individuals to contribute content on online platforms. Contextualizing this in our work, popularity shock can be viewed as a signal of increasing reputation and might prompt users to continue contributing to the platform. Though these theories were proposed some time ago, rigorous empirical evaluation/validation of these theories in the context of popularity on online social media platforms have not yet been conducted.

In this paper, we study how do users' behavior changes after a popularity shock in terms of (a) frequency of posting, (b) the content, itself and (c) how long do they continue with their altered behavior. We first characterize what should be considered as a popularity shock and develop a method to identify popularity shocks from a user timeline. Using popularity shock as an intervention, we use causal inference techniques to examine the change in behavior from pre-and-post popularity shock. Next, we study the change in the content posted by users under the effect of popularity shock. We leverage document embeddings [64] to model the posted content mathematically. Finally, we investigate the expected duration for shock's effect and its dependence on other factors using survival analysis techniques.

Data and Code: We released the anonymized version of our data available at:
`https://precog.iiit.ac.in/research/effect-popularity/`

2.2 Related Work

Since our work is related to users' response towards increased attention, our related work flows from three main directions - (a) Effect of social feedback, (b) Attention Shocks and (c) Popularity/ Virality Prediction.

Effect of social feedback: Positive reinforcement or feedback has been a popular area of study among social scientists [103, 95, 104, 6, 79]. [103] demonstrated through experiments on around 60 children through a bowling game that positive reinforcement led to improvement in altruistic behavior in children, while punishment led to the opposite. This framework has been studied extensively in various settings, such as effect of positive feedback on promoting

safe behaviours in housekeeping [104] and effect on compliance following transgression [79] as well as simulating motivations and future play of a brain training game [12]. In the domain of online world, however opposite effect has been observed in the case of low quality comments [16], where it was observed that negative feedback prompted users to continue with writing low quality comments on news articles. Further, [15] how the community perception of helpfulness of online reviews, influences consumer purchase decisions, and how this helpfulness vote is itself determined by evaluations of the same product by the community [22, 110]. Similar study has also been conducted for the effect of social feedback on weight loss community [20].

Though there has been a lot of studies discussing social feedback, however very few have tried to characterize how do users or actors in turn respond to extremely high and sudden feedback in data-oriented fashion on a large-scale data.

Attention Shocks: Attention shocks are characterized as sudden attention being drawn towards a specific entity (any author/artefact on social media platform). Examples include, death of a celebrity leading to increased attention towards the celebrity's wikipedia page [132]. [73] use the lens of organisation change to study the dynamics of change in behaviour of contributors of a GitHub repository experiencing increased attention as a result of being listed on the trending page. On Wikipedia, [132] observe increased participation of new comers and study collaborator dynamics on pages in times of shock detected through Google Trends, while [131] look into the changes in collaborative behaviour of editors due to shock resulting from imposition of censorship in mainland China. Other works like [52] study similar changes in case of breaking news articles on Wikipedia. [60] analyse shocks in form of sudden bursts of negative attention towards controversial events called 'firestorms', and use Twitter data to characterize the size and longevity of these firestorms.

Other works study the effect on online network structures under shocks. [53] suggest the formation of complex but temporary collaboration networks of users during increased editing activity on Wikipedia page of a diseased person and study their dynamics. Further, [51] introduce a method of capturing collaboration structure of co-authors of a Wikipedia articles and highlight the difference between such networks for breaking news articles, as compared to traditional ones based on pre-existing knowledge.

Though attention shocks have been studied on online social media platform, to the best of our knowledge, our work is the first attempt to study the behaviour of users whose posts goes viral (i.e. the user who gets the shock). A minor characteristic that differs us from other studies is that we are looking at shock as a sudden virality of the post, and the virality of the post is mostly algorithm-driven (i.e. probably a mixture of recommendation algorithm and "rich-gets-richer" theory). In comparison, other studies looked at shock which was more exogenous i.e. appearance on GitHub trending page or death of a celebrity. Lastly, there are inherent differences in nature of platforms being studied. While Github and Wikipedia are collaborative platforms where users are often driven by non-monetary motivations such as reputation and collective identity [73], users on such content sharing platforms are driven by

monetary causes and for self-satisfaction. Thus there is clear distinction in intent of use, due to which we can expect difference in user behaviour as well.

Though it is not highly aligned with our work, however there has been significant amount of work done for predicting if a post is going to get popular or not, and hence we mention about some of the efforts done to solve that problem.

Content Virality Most work in this domain is focused on predicting and characterising virality of online content. [28] understands popularity trends for online user generated content (UGC) in the form of online videos, and proposes a prediction model based on extremely random ensemble tree to predict the popularity trends for Youtube videos. The SEISMIC model proposed by [133] predicts the final number of reshares a post will recieve based on the past history. The problem is modelled as predicting the final size of an information cascade and performance is validated on a month of Twitter data. Other models like [127], [80] have tackled the problem of virality prediction on Twitter and Flickr respectively.

Other works are inclined more towards characterizing virality and viral content. [129] studies the virality and diffusion of memes on online networks. [76] seeks to identify features in posts which are related to its popularity using a multi-modal approach. [30] aims to characterize and understand popularity growth of videos, and what kinds of mechanisms contribute towards popularity. The work also mentions presence of sudden bursts of popularity on top listed videos.

Table 2.1: Number of unique users for each category (arranged in alphabetical order of Category).

Category	**Hashtags**	**Unique Users**
Animals	cats, dogs, pets	1666
Beauty/Makeup	beauty, makeup, naturalbeauty, skincare	3052
Craft/DIY	5_min_craft, craftchallenge, diycraft, easycraft	1128
Dance	dance, dancechallenge, dancekpop	2144
Education	careergoals, education, learning, mindpower	2429
Entertainment	entertainment	449
Fitness	fitness, fitnessgoals, gym, weightloss, workout	3911
Food	food, foodislove, foodrecipe, healthyfood, myrecipe	2815
Funny	comedy, funny, meme	2632
Health	wellness	558
Motivational	advice, inspirational, lifehacks	2146
Music	hiphop, music	1323
Pranks	prank	667
Sports	cricket, football, sports, tennis	2341

2.3 Theory and Research Questions

Kollock [57] hypothesized that there are three significant reasons for users to keep on contributing to the social community - (a) anticipated reciprocity; user is generally motivated to contribute or stay as an active participant in online communities in the expectation that the user will receive helpful information when they are in need, (b) sense of efficacy; the users might contribute information because they are rewarded with the sense that they contributed something to the community [9]. The efficacy can also result in the self-belief that they have a high impact on the community, hence providing the validation of their self-image as an efficacious person, and (c) Reputation; most users want recognition for their contributions or their efforts. As quantified by the number of unique impressions of their content, popularity validates their content. This can be seen as an increase in reputation for the user based on the high number of people that follow or subscribe to them. On the lines of Kollock, we hypothesize that receiving a popularity shock (i.e., increase in reputation) will prompt users to increase their activity on online social media platforms. [1] Therefore, we ask the following question:

RQ1. [Engagement Response to Popularity] *Do users increase their posting behavior after receiving popularity shock?*

Another social theory framework that fits very well with our setting is that of *operant conditioning* [111]. Skinner theorized that the reward for action leads the agent to keep on performing the same action in anticipation of reward, and a punishment hinders the user's propensity to take that action. Again, operationalizing reward as the popularity shock, we can hypothesize that users who received popularity shock will continue with the same behavior that earned them the reward even in our setting. This brings us to the following research question:

RQ2. [Content Response to Popularity] *Do users alter their content post receiving popularity shock?*

In network science, the transition of network states and dynamics due to an external event has been a topic of interest [132, 73, 53]. [74] argue that some of the network transitions, and along with it changes in user behavior in these networks, are more permanent. Moreover, some studies argue that networks bounce back after the event, and normal communication ensues [60]. In our setting, we were interested in understanding how long the popularity shock lasts.

RQ3. [Longevity of Effect] *How long do the effects of popularity shock last?*

For users who receive the popularity shock, it is imperative to understand what users can do or how they should maintain their activity that can prolong the shock's effect. Therefore we ask the following question:

[1] In this work, we discovered that the popularity shocks were positive, analysis can be done if this popularity instead was negative too.

RQ4. [Sustained Shock Effect] *What type of activity characterizes long-term sustainability of effects of popularity shock?*

2.4 Data Collection

Background: We collect data from popular multimedia sharing social media platform.[2] On the platform, users can post multimedia content (images/videos) along with an associated caption. Depending upon the privacy setting of the post and the user's profile, other users can view their content and engage with the content using platform-provided mechanisms such as liking the content, commenting on the post, or resharing the post. By liking, a user can express their positive response or acknowledgment, sharing works to amplify the reach of content, and viewers can also express their opinions in the form of comments. Like all other social networking platforms, the social platform understudy also provides functionality that allows users to 'follow' other users on the platform. Besides this, the platform can also grant a special 'verified' status to specific users based on their strong influence on the platform or in the real world. Though we study a specific platform, we believe that a similar methodology can be applied to any social media platform with similar mechanisms in place. A cross-platform study on measuring this behavior and ensuring generalizability is one of the promising future directions of this work.

Data Collection: We identify 14 generic categories related to commonly posted content on the platform. From the list of these 14 categories, we curated a list of 43 popular hashtags. The hashtag selection was made keeping the goal of generalization in mind, and hence no hashtags related to specific entities (e.g., #ronaldocr7) were considered. The selected categories and hashtags are described in Table 2.1. Approximately $4,000$ posts per hashtag were collected, coming from $21,224$ unique users. Next, we collect posts liked by these users and add the authors of the posts to our dataset to minimize any sampling bias due to the collection strategy(which might be due to bias in the platform's search functionality) [3]. Finally, we had a total of $33,490$ users. We collected the entire timeline of these users and filtered out users who had less than 200 posts in their entire lifetime to ensure we had substantial data for our analysis.

Following the filtering, our final dataset contains a total of $30,969$ users. We describe the data statistics in Table 2.2 along with distribution of number of posts across users in Figure 2.2.

For each post, we collected the following details of the post - (a) post id, unique identifier for the post, (b) timestamp of when the post was published, (c) caption of the post, (d) number of views the post received, (e) number of likes the post received, (f) number of times the post was reshared, (g) number of comments the post received and (h) user information - all key statistics such as name, bio, etc. of the user who created the post.

[2] name of the platform suppressed to retain anonymity and non-public API access.

[3] Data collection was done when the first and second authors were students at their respective institutes

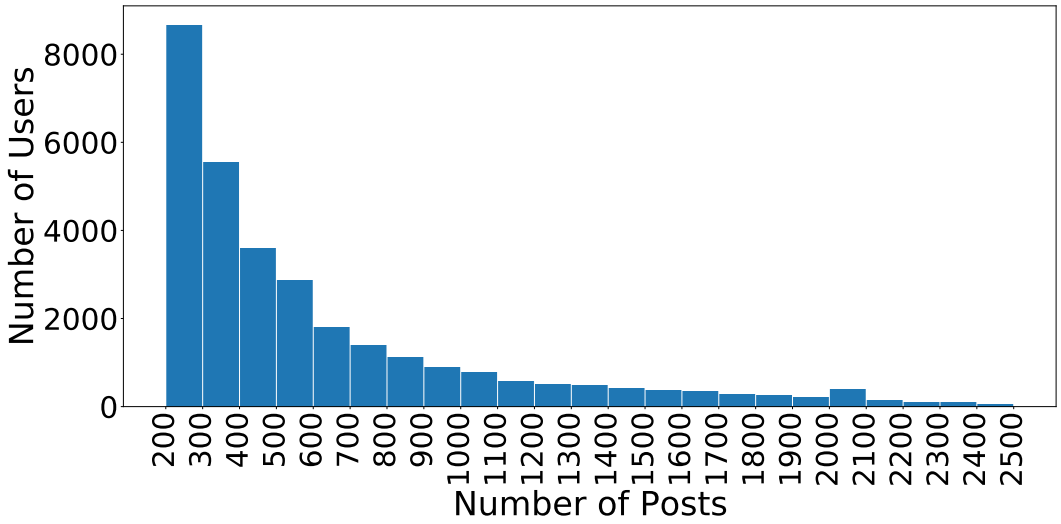


Figure 2.2: Distribution of Users' Total Posts (Follows Power Law).

Table 2.2: Dataset Details.

Number of Users	30,969
Number of Posts	18,911,417
Timestamp of First Post	7^{th} Jan 2015
Timestamp of Last Post	31^{st} Dec 2021

2.5 Detecting Popularity Shocks

To answer any of the research questions mentioned above, we first need an algorithm that can identify popularity shocks from a user's timeline. Before going into the details of the algorithm, we describe the assumptions we made to define popularity shock.

- We use the number of views as a proxy for popularity. Views give a more objective metric of the reach or engagement as it is implicit, unlike other metrics such as the number of likes, shares, or comments which require explicit action from the audience.
- A user might receive multiple popularity shocks throughout their career. However, we only study effects due to the chronologically first shock the users receive. We do not consider later shocks as the user would have already experienced some popularity until that point. In this paper, we want to characterize the effect of the first popularity shock when the sudden growth in popularity is unexpected for the user.

A desired shock detection algorithm should detect a sudden percentage increase in views of the user, we should also account for absolute thresholds to avoid false positives caused by the base effect. The first natural candidates for the task are time-series anomaly detection algorithms like Z-score [14] or Facebook's Prophet [117]. However, these algorithms consider

time-series signals in isolation and do not account for global thresholds. To curb this, we also experiment with a custom algorithm as presented in Algorithm 1.

We preprocess the timeline by binning the posts, where each bin is a period of consecutive D (bin size) days. For each user, we iterate over the bins in chronological order (Line 6). We maintain a running average of views of all the bins encountered so far (Line 13). Once we have processed the bin (i.e., no more posts need to be counted for that bin), we compute the ratio of views of the bin to the running average of bins before it. Note that we ignore bins with no posts while computing the running average. This ratio needs to be higher than a ratio threshold θ for it to be considered a shock candidate. To account for the cases where the running average is very low, we also consider the difference between current views and the running average, which needs to be greater than the base threshold η. Therefore, the first bin satisfying these two conditions is classified as the popularity shock for the user. If no point satisfies these conditions, we consider the user is without a popularity shock.

Ideally, keeping consistent with our shock assumptions, we want to capture the first post at which user perceives they might have gotten popular. To evaluate our detection algorithm, we conduct a verification experiment. We solicited annotations from long-term social media users, who were asked to independently look at the view timeline of 100 users and mark what they deem as the first instance a user would have felt popularity shock. The annotators had a Fleiss' Kappa score [31] of 0.60, which indicates moderate agreement [63]. Each sample was annotated by 3 annotators, and a clear majority was received in 93 instances out of 100. We compared the efficacy of our proposed approach with baselines of z-score and Prophet algorithm using the ground truth set. Predictions were obtained across a range of hyper-parameters for all algorithms, best achieved results are shown in Table 2.3. Our proposed shock detection algorithm performs the best and is used to detect popularity shocks in our further experiments.

Table 2.3: Shock detection accuracy against the manually annotated ground truth. Proposed algorithm outperforms other baselines.

Algorithm	Accuracy
Z-Score	23.6%
Prophet	42.5%
Proposed	66.6%

The percentage of users we can discover having a popularity shock with different values of θ and η is presented in Figure 2.3. Note that, we report results with hyper-parameters $D = 1$, $\theta = 50$, $\eta = 1.5$M, unless specified. However, we experimented with multiple values of θ and η, and results stay consistent over reasonable values of these thresholds.

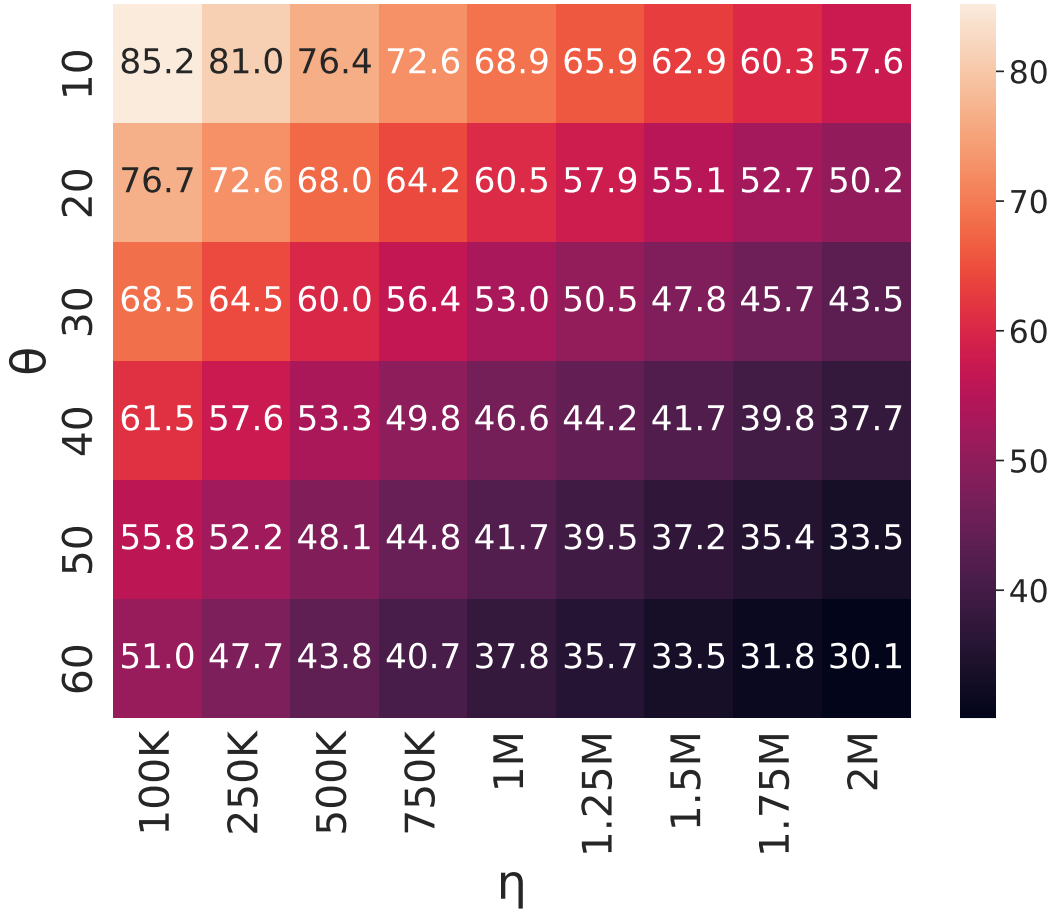


Figure 2.3: Heatmap representing percentage of users detected with a shock for different values of θ and η for $D = 1$. θ is the minimum ratio of views in the bin to the running average, while η is the minimum difference between the two, for detecting shocks.

2.6 Effect of Popularity

RQ1 seeks to quantify the change in posting frequency of a user due to the shock received. We do this using a causal inference technique called Regression Discontinuity in Time (RDiT) [41].

Regression Discontinuity Design (RDD): Introduced by [119], RDD is a quasi-experimental technique to measure the effects of a treatment or intervention. The population receives the treatment having the value of running variable X above a certain threshold known as the 'cut-off' point, and data is checked for any jumps or discontinuities in the outcome variable Y around the cut-off. Previously, RDD has been widely used in fields such as Economics [66] and Psychology [17]. Specifically, on social media studies, RDD has been analyzed previously to quantify the effect of obtaining a GitHub badge on users' posting frequency [92], on the effect of the introduction of Facebook "People you may know" feature [74], and also on the effect of averaging rounding stars on Yelp [68].

Algorithm 1 Shock Detection Algorithm

function DETECTSHOCK(posts, D, θ, η)
    bins $\leftarrow$ *bin_data*(posts, bin_size=D)
    shock $\leftarrow -1$
    n $\leftarrow$ *length*(bins) ▷ Total number of bins
    run_avg = *views*(bins[1])
    for i in 2 to n **do**
        ratio $\leftarrow$ *views*(bins[i])/run_avg
        diff $\leftarrow$ *views*(bins[i]) $-$ run_avg
        if ratio $\geq \theta$ and diff $\geq \eta$ **then**
            shock $\leftarrow$ i
            break ▷ break at the earliest shock
        end if
        run_avg $\leftarrow$ *mean*(*views*(bins[1:i]))
    end for
    return shock ▷ if shock is -1, no shock found
end function

Acknowledging that time being the running variable might cause some of the assumptions of traditional RDD not to hold, we use a variation of the RDD framework called **Regression Discontinuity in Time (RDiT)** proposed in [41], in which time is the running variable and a fixed point in time is taken as the threshold. RDiT conceptually differs from the regular RDD on the following fronts:

- While RDiT aligns with the 'discontinuity at cut-off' interpretation of RDD, the 'local randomization' interpretation may not hold as the time assignment can not be taken as entirely random around the cut-off.
- Unlike RDD, sample size can not be grown arbitrarily with smaller bandwidths. Due to this, data points far from the cut-off need to be included, which can introduce biases due to changes in unobserved confounders over time.
- Including covariates becomes far more critical to control biases since the assignment of treatment and control groups is not entirely random around the cut-off.

Our methodology: To model our problem using RDiT, we define our running or forcing variable X as the bin index (signifying time) and outcome variable Y as the number of posts done by the user in the bin X. The shock bin is assigned index 0; subsequently, index $+i$ denotes the i^{th} bin after the shock, while the index $-i$ denotes i^{th} bin preceding the shock. The cut-off point c is $X = 0$, where the shock occurs. Then, treatment group is defined as $\{(X_i, Y_i)\ s.t.\ X_i > 0\}$ and control group as $\{(X_i, Y_i)\ s.t.\ X_i < 0\}$. We also control for the following covariates in our regression design:

- **Intensity of shock:** To account for variation in treatment, we control for the intensity of shock obtained in the preceding bin. The intensity is the value of the *ratio* variable for the bin as in Algorithm 1. We take the logarithm of this variable.

- **Age of User:** As receiving a popularity shock at different stages of users' online life might have different effects. We control for the number of days since the user's first post.

We then fit models separately on the two groups using regression. We only use W bins before and after the shock bin to fit the lines to avoid any effects of future shocks. On obtaining the equations of the two lines, their values at the cut-off point are predicted, which are used to calculate the discontinuity at the cut-off. Formally, let $Y_{t,0}$ and $Y_{c,0}$ be the values at the cut-off for the treatment and control lines respectively, then discontinuity at the shock d is given by $d = Y_{t,0} - Y_{c,0}$. From the equation, it can be seen that a positive d corresponds to an increase in the frequency of posting after the shock as compared to before and vice versa.

2.6.1 Effect on Posting Frequency

We tried to estimate the effect of popularity shock on the posting frequency of user post-shock using RDiT. We quantified the intervention to occur at the time-point where we detected the popularity shock. Further, we count the total number of views that the user received each day before and after the shock. Note that this corresponds to setting $D = 1$ in Algorithm 1.[4] In Figure 2.1b, we visualize the effect on posting frequency. The x-axis clearly shows the time before and after the shock. To aggregate the effect across all users, we compute the number of posts done by the user each day subtracted by the average number of posts done by the user in the past 15 days (this is done to maintain a consistent scale across users). Then, the average is taken across all users(including covariates) and curves as fit. The vertical dashed line shows the day on which popularity shock was observed. As mentioned above, we fit two linear regression models.[5] The first model is for the average number of posts done before the popularity shock, and the second one is for the average number of posts after the popularity shock. We see a significant difference between the intercept and the slope for both the regression models. The discontinuity at shock (d) estimates how users are changing their posting behavior pre- and post-shock. This is measured as the difference of the predicted number of posts done at the shock by the two regression models (intercept of the second model - intercept of the first model). We note that for all values of W, we observe positive discontinuity, implying a positive effect on the number of posts made by the user after receiving popularity shock. Both of these slopes are significantly different and hint

[4]Important to note that here, we also experimented with various values of D, θ and η and achieve similar results.

[5]We also experimented with higher order polynomial regression models, and results were consistent. Although we do observe overfitting in some cases.

towards a significant effect due to shock. Looking at the regression fits and the magnitude of discontinuity, we make the following observation:

Observation 2.1 (Increased Posting) *Users increase their posting behavior post shock.*

Observation 2.2 (Short-Term Gains) *Though users increase their posting behavior post shock, it also quickly decays off, as time progresses.*

Note that while the trend of the fit of the model pre-shock is positive and post-shock is negative - this could be due to the sensitivity towards our shock detection algorithm. Our shock detection algorithm works by binning the posts and classifying if a particular bin is a shock bin or not, and also, the algorithm takes into account total views rather than the average number of views. Therefore, users might be posting a high number of posts that were getting a sizable number of views (lesser than our threshold) until eventually tipping on the next bin and satisfying our threshold.

2.6.2 Significance of Result

We perform following checks as mentioned by [41]. 1) We control for observable confounders to remove biases and account for variation in treatment. 2) We perform a Placebo Test to ensure no discontinuity at points where there should not be any. [45] suggests checking for any discontinuities at the median values of the running variable for the sub-samples corresponding to either side of the cut-off and using standard errors to test for no discontinuity. We do this test only for the sub-sample below the cut-off, as the points above our cut-off may have discontinuities due to potential future shocks. Say the shock occurs at the s^{th} bin from the start, then we check for any discontinuity at $\frac{s}{2}^{th}$ bin. We observe significantly less discontinuity and overlap between 99% CI intervals, implying no observable discontinuity. 3) We check for robustness of our results towards window size and polynomial order. 4) We fit regression lines without controlling for covariates and observe similar results, indicating no time-varying treatment effects.
Note that, as suggested in [41], the McCrary density test [77] is not valid when time is the forcing variable. However, we argue that there is no manipulation in our case as users' can not preempt an imminent shock due to lack of knowledge of platform recommendation algorithm and the large magnitude of our shocks (50x more views with 1.5M difference).

2.6.3 Effect on Posted Content

In **RQ2**, we aim to determine if users alter the content they post after receiving a popularity shock. We characterize the content by using the posts' captions. The posts' captions can be noisy, so we take appropriate steps to develop a consistent representation from the captions. First, we preprocess the hashtags present in the caption by removing the '#' symbol from every hashtag and then use wordsegment[6] library to segment these hashtags into separate

[6]http://www.grantjenks.com/docs/wordsegment/

Table 2.4: Results showing similarity of content for before and after the shock to the shock($^{***}p < 0.001$).

Time Period	All Users	
	Sim(Pre, Shock)	Sim(Post, Shock)
7	0.625 ± 0.22	$0.714 \pm 0.17^{***}$
30	0.656 ± 0.21	$0.699 \pm 0.20^{***}$
	High Discontinuity Users	
7	0.645 ± 0.24	$0.730 \pm 0.20^{***}$
30	0.670 ± 0.21	$0.732 \pm 0.19^{***}$

words in order to extract their semantic meaning. Following this, we compute the similarity between the content posted in two time periods (set of bins). We represent the captions of all the posts done in that bin duration using a single feature vector and then measure their similarity. We use document embeddings to come up with the representation. We convert every post into a single vector using the document embedding of its caption. We leverage doc2vec [64] to generate embeddings.

Subsequently, we obtain a single vector representation for a time period by averaging the document embedding vectors corresponding to a set of posts from that temporal bin. We use cosine similarity to compare vectors formed using document embeddings. Cosine Similarity yields a score between 0 and 1, with 1 representing the same vectors. With the above experimental framework, we compare content posted in the shock bin with that of W bins just before and after the shock to capture the change around the shock. We also perform the analysis for users whose discontinuity in posting frequency lied in Top 25 percent. Based on the results in Table 2.4, we make the following observations:

Observation 2.3 (Post Shock Similarity) *Users, post-shock generate more similar content to the shock inducing posts.*

Observation 2.4 (High Discontinuity Similarity) *Users who increase their positing frequency more, also tend to stay more closer content-wise to the shock related posts*

We can observe from Table 2.4 that similarity of *doc2vec* embeddings between post-shock and shock is significantly higher than similarity between pre-shock and shock. We use significance test and obtain $p < 0.05$ to show that these two values over all users is significantly different.

2.7 Sustainability of Popularity

In both **RQ 3** and **RQ 4**, we try to answer the questions related to the sustainability of the popularity shock. For both of the questions, we leverage *survival analysis* [83]. Survival

Analysis is a popular multivariate event history modeling technique that focuses on estimating the average hazard rate of an event under consideration at a given time and also corresponding relative strength of the effect of different factors on this hazard rate, where hazard rate can be defined by $h(t) = \frac{P(T<t+\delta|T\geq t}{\delta t}$. Cox proportional hazard model [18] can be used to estimate this probability and the coefficients of the regression $h(t,X) = \theta(t)\exp(\beta^T X)$ using partial likelihood, without making any assumptions about the baseline hazard rate.

Our observation period for a particular user starts from the bin where the shock occurs. We define our event of interest as the point in time post the shock where there is no difference in activity level compared to pre-shock level. Specifically, we rely on the number of views to compare post-shock and pre-shock levels. We say that the increased response due to shock has faded away if we discover B consecutive bins with the number of views less than K. We set K as the 10% of the views obtained in the shock bin. We set the value of B as 3.

2.7.1 RQ3: Longevity of Shock Effect

In RQ3, we study how long the effect of a popularity shock lasts. We plot in Figure 2.1c, the survival curve for users to demonstrate the longevity of effect on shocks. From the curve, we observe that the effect of shock dies down rather quickly for most users. For 50% of the users the effect fades away in the first 5 days itself, while it ends for 90% of the users within 39 days of the shock. This implies that it is extremely difficult to maintain response levels observed during the shock for an extended period.

Observation 2.5 (Shock Longevity) *Popularity shocks are short-lived. The increased response received by users goes down to pre-shock level very quickly after the shock.*

2.7.2 RQ4: Sustaining Shock Effect

In RQ 4, we model the factors on which the longevity of shock effect depends as well as the effect and extent of the dependence. To do this, we build on existing survival model, and use Cox Proportional Hazards regression model [18] to quantify the effect of different factors on survival.

Factors affecting survival: We are specifically interested in understanding what a user can do to prolong the effect of popularity shock. We hypothesize the following factors:

1. *Posting frequency*: The frequency of posting represents how eager a user is to create and post more content after the popularity shock. It can be hypothesized that high posting frequency could indicate users trying to be more active on the platform and trying to engage highly with the new audience that the user has got access to. We operationalize this by the total number of posts a user does in a bin.

2. *Similarity in Consecutive Posts*: The change or variation in the content that users post could be indicative of how versatile the user is in adapting their content to the needs of their audience. A user might have got popular due to a specific type of content and keep

Table 2.5: Dependence of Shock Effect survival on other variables using Cox Regression ($^{***}p < 0.001$).

Covariate	**HR (St Err)**	**LR Chisq**
Avg. Likes	0.90 (0.01)***	292.43***
Shock Intensity	1.13 (0.03)***	80.59***
Posting Frequency	0.86 (0.01)***	1047.9***
Similarity between consecutive posts	6.54 (0.03)***	2734.31 ***
Similarity of posts with shock post	0.38 (0.04)***	37.57 ***

posting it in the hope of a similar response. However, this may lead to repetitiveness in content, and the audience might lose interest. Our analysis operationalizes this by the average cosine similarity between all posts in consecutive bins.

3. *Similarity with the shock content*: The similarity between the shock-related content and the current content is an indicator of how much the user has digressed from the content, which leads to their popularity. Viewers often start associating users with a specific type of content, and thus deviating too much from that may cause disengagement from their audience. We model this as the average cosine similarity of content posted in a bin with the shock content.

Though these are the factors that we are interested in, we also control for the following variables, which could affect the longevity of the effect.

- *Effect of feedback*: The amount of feedback received by a user on the posts user created after popularity shock is indicative of the engagement levels of the user's audience. We measure this by introducing three variables - (a) Number of likes, (b) Number of shares, and (c) Number of comments. Since these variables are highly correlated, we only use the average number of likes in the regression model.

- *Intensity of shock*: Another factor that needs to be controlled as to what was the magnitude of the shock. Higher the intensity of the shock, higher will be the survival chance for it.

We report the results of Cox proportional hazard regression model in Table 2.5.

Observation 2.6 (Constant Posting) *Maintaining high posting frequency helps keep retaining the long-term effect.*

Observation 2.7 (Similarity in Content) *Users deviating away from the content which got them to the shock have shorter survival times of shock effect, at the same time having high*

similarity in consecutive posts can lead to repetitiveness which again causes the survival to go down.

Observation 2.8 (Engagement) *On audience side, high engagement from audience helps maintain the effect of popularity shocks.*

2.8 Discussion and Implications

2.8.1 Research Questions

In this paper, we focused our analysis on popularity shocks. We started with four research questions related to the effects of popularity shocks, longevity, and sustainability of the shock. Specifically, **RQ1** tries to study the effect of popularity shock on users posting frequency. From the RDiT results, we discover that users increase their posting frequency after the shock compared to before. However, as time passes, the posting frequency starts to decrease. **RQ2** is concerned with analyzing how does a user changes the content that they post after popularity shock. We find that not only do users alter their content after the shock, the post-shock content is also more similar to the content which leads to the shock, as compared to before. Thus, we conclude that popularity shocks indeed induce a behavior change in users who experience them. We are interested in understanding the longevity of the popularity shock, and hence we ask the **RQ3**. We used survival analysis to answer this question. We observe that most shocks are short-lived, i.e., the shocks reduce to 10% of their shock intensity within 5 days for 50% users. For **RQ4**, we were interested in knowing the factors that enhance the sustainability of popularity shock effects. We discover that repeatedly posting the same content as well as deviating away from the shock content cause low shock survival. Finally, high posting frequency and high response received from the users lead to more prolonged shock effect survival.

It is also worth discussing that a popularity shock or virality may not always occur in a positive connotation. Such shock can also indicate hate or networked harassment (i.e. negative attention) towards the creator [67]. Similarly, increased content posting frequency can be attributed to the author apologizing, explanation, or clarifications. Such hateful phenomenons can adversely affect the mental health of the creator [94] and cause instability in the community [10]. Though our work is centered only on positive popularity shocks, a potential extension to our work can be to categorize shocks into positive or negative and analyze their effect on the creator's behavior.

2.8.2 Implications

Our paper provides numerous insights and observations into phenomena of popularity shocks. These insights form the basis for several implications for all three - (a) advertisers, (b) platform designers, and the (c) users.

Advertisers, or brands can adjust their marketing campaigns by understanding which users are behaving in a particular fashion that will lead to lasting popularity levels. They can also use topical information to identify if popular users identify more with their brand's content or not.

Platforms can utilize the insights from the study to devise algorithms for their trending pages. As popularity shock is found to increase users' engagement with the platform, enhancing attention towards dormant users can cause them to resume to increase their activity. Our content similarity results also show that such shocks can cause homogenization of content on the platform.

Users can learn the behaviors which lead to sustaining the effect of popularity shock. This can help them keep their increased engagement and benefit from the shock for a longer duration.

2.8.3 Threats to Validity

Like any quantitative study, our work is subject to multiple threats to validity. In this section, we attempt to list biases, data issues, and threats to the validity of our study by following the framework proposed by [93]. First, our work is based on a single social platform, and though it works and leverages features available on multiple social platforms, similar results do not have to hold. One possible point of differentiation could be that each platform has a different recommendation algorithm for recommending content to its users. However, the effect of recommendation algorithms on our results should be minimal since we study the effect of receiving a popularity shock by the user whereas, the recommendation algorithms primarily determines who and how big of a shock user will get. Our data can also suffer from representativeness - we use just a limited set of users who posted using a limited set of hashtags. This data representation could be significantly different from the general population on the platform. Another data issue that theoretically casts clouds on the analysis is that the number of views, likes, and comments are retrospective, i.e., they are not computed in real-time while they are the numbers on the platform at the time of data collection. Though we believe the practical effect on our results is limited since the majority of impressions on social media posts are received soon after posting [133]. For further validation, we tracked daily view counts of $1,374$ randomly sampled posts for the first 10 days after posting and found that 70% of total views were received in the first 2 days. Additionally, we did perform two analyses - regression discontinuity and survival analysis. We ensured that our data and modeling choices hold the assumptions, but there might be some unobserved confounders that we might not have considered. Finally, our statistical modeling required multiple parameters related to the operationalization of theories in sociology literature. Some of these parameters might not be capturing the factors that we intended to capture or that the theories proposed.

This work forms the basis for various future works related to popularity shocks. First of all, the work can be extended to a more generalized population and more social media platforms. Similarly, extending to different users could also open the potential to study the

effect of user personality or user type on how they respond to popularity shocks. Another significant improvement in this work could be by leveraging matching techniques to match users who got popular with similar content with users who did not get popular and then record average responses. This was not possible in our current work due to multiple reasons - (a) limited data and (b) the presence of too many confounders to create a propensity model for popularity prediction.

2.9 Conclusion

We performed a large-scale analysis of the effect of popularity shocks on users. Grounded in operant conditioning and increased sense of reputation, our results confirm the extent to which popularity shock leads users to post more and modify their future content to be more similar to the content that made them famous. Similarly, on analyzing the longevity of this shock, we discovered the short-lived nature of the shocks and the effects of various posting behaviors on shock longevity. We also provide factors that users could leverage for sustaining increased engagement post-popularity shock.

Chapter 3

Effect of Feedback on Drug Consumption Disclosures

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 so again. To this end, we collect data from 10 drug-related subreddits. First, we build a deep learning model to classify user-generated content as indicative of drug consumption offline or not. We 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 in drugs.

3.1 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 [90]. The US president declared the drug crisis as a national public health emergency in 2017.[2]

[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

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 [71]. 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 [44], forecasting drug overdose [81], transition into drug addiction [71], patterns of drug use and consumption methods [8], and geospatial patterns in drug use [7]. 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 [19], physical activity [2], smoking and drinking relapses [116], quality of user-generated content [16] and involvement in open-source projects [123].

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 consumption or not, to quantify the extent of drug consumption activities. Further, grounded in *Primacy Effect* [4] and *Operant Conditioning Theory* [112], we use propensity score matching [115] 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 indulged in 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

[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.

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 social media content 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 lead to harm reduction.

Data and Code: Subreddit data is available via Pushshift API. [5] Our annotated dataset, user study responses, and modeling code is available at *https://doi.org/10.5281/zenodo.6041837*.

3.2 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 [35]. Leary et al. [65] showed that individuals indulge in voluntary risk-taking activities like consumption of drugs, distracted driving, unprotected sex to improve impression among peers. Hogan [43] 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 [4]. For e.g., people's impression of an individual is dependent

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

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

on the first traits they encounter [4]; probability of recalling initial items in a list is higher [86]; people have a more vivid memory of their first romantic encounter, achievements, and even losses [27]. The primacy effect can cause *anchoring bias*, leading to skewed decisions relaying heavily on the initial information [121]. Building on these theories, [107] 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 [120]. Based on this principle, Skinner et al. proposed *Operant Conditioning* [112]. 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 [19], increased social media interactions lead to higher steps in activity tracking apps [2] and community feedback affects the quality of future posts [16]. 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 representations of text. Mikolov et al. [82] 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. [124] proposed an improved model architecture called Transformers based on self-attention [106] to learn contextually aware dense text representations. Transformer-based large pre-trained models [25, 70] 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?*

3.3 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 online 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. [71] 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. [8] used Word2Vec [82] 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. [7] 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, [44] used deep learning ensemble models to detect drug abuse in tweets and [81] used community attentive neural networks to forecast drug overdoes using information about crime dynamics.

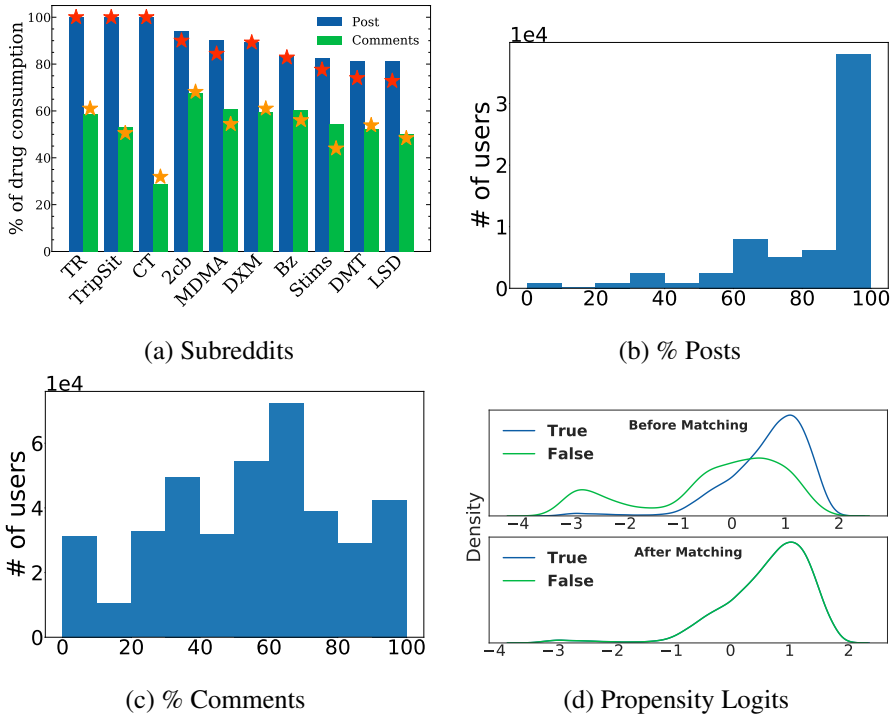
Causal inference using online data: Traditionally, researchers have established randomized controlled trials to establish causations. However, due to logistically and ethical concerns, such trials are not always feasible; e.g., it is not ethical and legal to make subjects consume illicit substances to 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 [101].

[19] showed that positive feedback from the online community could help users lose more weight. [2] studied data from an exercise logging application and found increased social connections on the platform caused higher physical activity in the offline world. [116] used survival regression to establish a causal relation between linguistic cues from user-generated content and smoking or drinking relapse. [55] 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. [23] unveiled the causal relation between user's vocabulary and suicidal tendency.

Online data have also shown effects in opposition to expectations, e.g., [16] 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 [65]. An increasing number of users are getting involved in dangerous social media challenges like the KiKi Challenge [5], the Salt and Ice Challenge [102], the Cinnamon Challenge [38], Tide Pod Challenge [87], and the

Fire Challenge [1]. [61] 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, [88, 59] 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 3.1: (a) Perecentage of drug consumption content across subreddits. Values derived from proposed model are indicated by bars, and $\star$ 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.

3.4 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 to the users called *score*.

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

Table 3.1: Statistics about the data collected.

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
Benzodiazepines	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
Currentlytripping	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

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 3.1), 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 3.1 provides a summary of statistics for each subreddit.

3.5 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,

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

[9] `https://www.reddit.com/r/help/comments/3tbuml/whats\_a\_flair/`

number of comments, and score have on their future participation in subreddit and drug consumption. Each question could be answered on a 5 point Likert scale, 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.

3.6 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 [71, 8, 7]. 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.

3.6.1 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-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...*

Annotators were also given a list of drugs *street* names and slangs used in drug-related subreddits to assist the annotation process [8]. Each sample was annotated by 3 annotators

independently. We obtain a Fleiss-kappa [31] agreement rate of 0.69, which signifies substantial agreement [63]. 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]

3.6.2 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) [56]. Input text for the Text-CNN model is vectorized using pre-trained Google News corpus Word2Vec embedding [82].

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 [124]. Leveraging that, we experimented with BERT, a model built using bidirectional transformers and pre-trained on masked language model, and next sentence predictions tasks [25]. We also built a

Table 3.2: Drug consumption classification performance.

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**

[10] `https://doi.org/10.5281/zenodo.6041837` (anonymous link for double blind review)

Table 3.3: Performance of proposed model across subreddits on test set. Sorted by Macro F1 score.

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
Currentlytripping	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
Benzodiazepines	82.35	86.24	71.17	74.09
MDMA	76.92	74.44	69.30	70.68

classifier based on RoBERTa [70], an optimized version of BERT. Table 3.2 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 3.2 provides performance of all models on test set. Our best model achieve a macro F1 score of **79.54**. Table 3.3 provides performance numbers of our best model across subreddits.

3.7 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 3.1(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.

Further, 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 3.1(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 3.1(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.

Observation 3.1 (Extent) *About 80% of total users in drug-related subreddits had consumed drugs in real life. This is inline with the data received via our user study.*

Observation 3.2 (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.*

3.8 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 [46].

In the potential outcome framework [89], 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 received. Conventionally, treatment is a binary variable (e.g., vaccine administered or not), and hence the assignment of treatment is trivial. However, 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. 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.

[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 refereed as Control group.

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.

Multiple recent social media causal inferences studies have used text-based models for propensity estimation [16, 114, 23, 55, 105]. Most of these studies use a combination of n-gram features and Logistic Regression to train the propensity model [54]. However, recently [128] 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 [128] findings we use pre-trained RoBERTa [70] 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.[12] 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 [101], and hence, the importance of classification performance is secondary [55]. Further, [128] 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: Each user in treatment group is matched with one user from control group with similar propensity score. Generally, given a propensity score p, matching is done on $logit(p)$ (Equation 3.1). A pair is considered as a good match, if difference of $logit(p)$ is less than a *caliper* value as defined in Equation 3.2 [40].

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{3.1}$$

$$caliper = 0.25 \times \sigma(logit(p)) \tag{3.2}$$

[12]Training parameters were similar to those presented in Section Detecting Drug Consumption Content. Training code is present in our code repository `https://doi.org/10.5281/zenodo.6041837`

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, frequency and feedback on past posts should also be balanced as confounders [16]. 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 [115] defined as: -

$$SMD = \frac{\overline{x}_{treatment} - \overline{x}_{control}}{\sqrt{\frac{\sigma^2_{treatment} + \sigma^2_{control}}{2}}} \tag{3.3}$$

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 [115].

In literature, where text propensity models are built on n-gram features, SMD balance check is conducted on n-gram vectors [19, 55, 105]. Since our propensity model is deep learning-based, we use feature vectors extracted from the last hidden layer of our model to conduct a balance check. 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 [115].

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:-

$$EATE = \frac{\sum_{i=1,j=1}^{N} \frac{(Y_i(T=1) - Y_j(T=0)) * 100}{Y_j(T=0)}}{N} \tag{3.4}$$

$EATE$ 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.

3.8.1 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. [13] We experiment with comment thresholds (θ) between 2 to 6.

[13] 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.

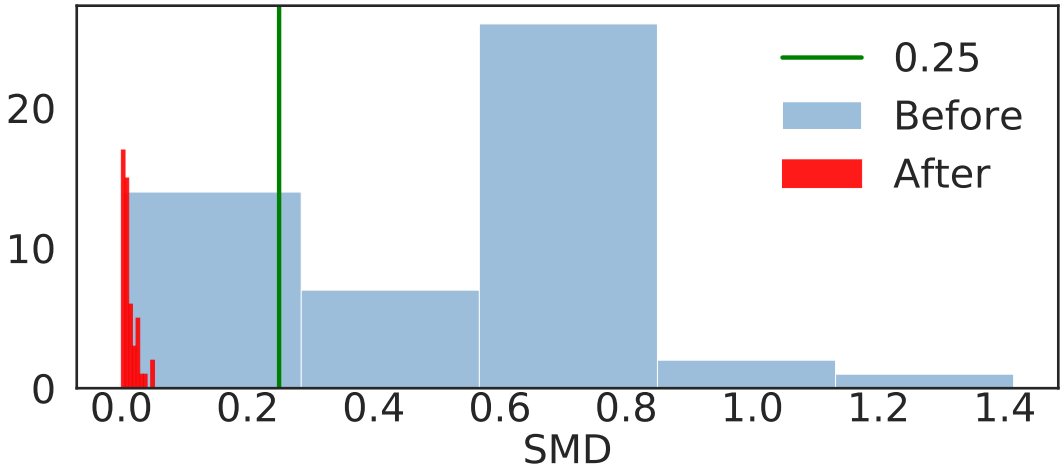


Figure 3.2: Matching quality for r/LSD, n_1, $\theta = 4$. Distribution of confounders' SMD before and after matching. After matching *SMD* for all confounders in ≤ 0.25 indicating good quality matching.

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 [75] and consistent across different treatment thresholds and subreddits. Table 3.4 shows $EATE$ of n_1 for all the subreddits calculated on $\theta = 4$. Figure 3.2 shows change in confounders *SMD* and Figure 3.1(d) changes in $logit(p)$ distributions before and after matching for *r/LSD* n_1, $\theta = 4$.

3.8.2 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 3.4 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 statically significant result that conflicts with our hypothesis.

3.8.3 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 3.4 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 ≈ 600 experiment configurations. Due to lack of space, it is not feasible to present results of all the configurations in the paper. 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://doi.org/10.5281/zenodo.6041837` (anonymous link for double blind review).

Observation 3.3 (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 [72]. Understanding the contrast between user opinion and statistical findings is vital to designing effective intervention and harm-reduction strategies.

Observation 3.4 (Effect Underestimation) *Users on drug-related subreddits tend to underestimate the effect community feedback has on their future engagement and drug consumption.*

3.9 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

Table 3.4: $EATE$ 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.

	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.

(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 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 statically. Discrepancies like this have been studied in psychology literature [72] 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.

3.9.1 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 pre-trained models from drug consumption classification to enable further research.

3.9.2 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.

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.

3.10 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, and the volume of such content is rising exponentially. 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 caused by such behavior.

Chapter 4

Social Re-Identification Assisted RTO Detection for E-Commerce

E-commerce features like easy cancellations, returns, and refunds can be exploited by bad actors or uninformed customers, leading to revenue loss for organization. One such problem faced by e-commerce platforms is Return To Origin (RTO), where the user cancels an order while it is in transit for delivery. In such a scenario platform faces logistics and opportunity costs. Traditionally, models trained on historical trends are used to predict the propensity of an order becoming RTO. Sociology literature has highlighted clear correlations between socio-economic indicators and users' tendency to exploit systems to gain financial advantage. Social media profiles have information about location, education, and profession which have been shown to be an estimator of socio-economic condition. We believe combining social media data with e-commerce information can lead to improvements in a variety of tasks like RTO, recommendation, fraud detection, and credit modeling. In our proposed system, we find the public social profile of an e-commerce user and extract socio-economic features. Internal data fused with extracted social features are used to train a RTO order detection model. Our system demonstrates a performance improvement in RTO detection of 3.1% and 19.9% on precision and recall, respectively. Our system directly impacts the bottom line revenue and shows the applicability of social re-identification in e-commerce.

4.1 Introduction

Over the last decade, e-commerce adoption has proliferated rapidly [91]. Such growth is fueled by convenience that e-commerce can provide over brick and mortar, e.g., large product selection, lower prices, same-day shipping, and hassle-free returns and cancellations. Though convenience features attract customers, they can sometimes cause significant business challenges; one such case is Return-to-Origin (RTO). RTO as depicted in Figure 4.1 is a scenario when a customer orders a product and then cancels while it is en route. RTO leads to two kinds of losses in a system:-

- **Logistical cost**: This is the cost of shipping the product till the point of cancellation in the supply chain and then returning it to the warehouse safely and restocking it.

- **Opportunity cost**: In the time while the product was ordered and canceled, this product unit became unavailable to order by another customer who would accept the delivery.

Though business accounts for potential revenue loss while offering functionality like RTO, an increased rate of RTO by uninformed customers or bad actors can cause unanticipated revenue losses totaling double-digit million dollars annually. Hence it becomes necessary to develop a real-time system that can predict the likelihood of the order being subjected to RTO at the time of checkout. Prediction of the model combined with other attributes like customer history, and available stock of the product can be used to initiate precautionary measures that can mitigate RTO risk. Naturally, the data used to build such a system would be, the historical pattern of RTOs at a user and product level. However, a system built on these features is limited in its capability, especially for new users and product categories.

Literature has shown that socio-economic attributes of customer can be an indicator to identify the likelihood of a person being involved in activities like electricity theft [96], false insurance claims [118], or mortgage fraud [13]. Public social media profiles can be used to estimate socio-economic features [62]. Adding features from social media profiles has shown improved results in a variety of tasks, e.g., identifying transaction fraud [48], the credibility of online information [39], hate speech [21], and propensity to participate in risk-taking activities [61, 65]. Grounded in the aforementioned literature, we hypothesize that enriching historical data with publicly available social data of a consumer will lead to a performance improvement in RTO prediction.

The first step for our experiments is to re-identify the social profiles of a given user. The problem of social re-identification is studied widely [47, 130, 108, 84, 85]. Though most literature relates to retrieving matches between two social media platforms with a notable exception of [36], our task is slightly different, where we need to match profiles between a social and an e-commerce platform.

In this paper, given an e-commerce user, we find the relevant public social profile and show that the fusion of social information with historical trend data improves the performance

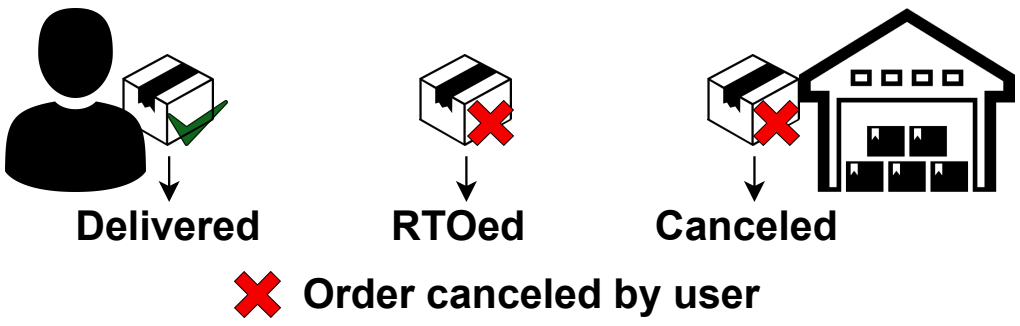


Figure 4.1: An order becomes Return to Origin (RTO) when the user cancels an order after it has been shipped from the source location.

of RTO prediction by 3.1% increase in precision and 19.9% increase in recall. Our work has direct implications for e-commerce platforms where a system like this can prevent loss of revenue. Additionally, our study demonstrates that combining social information with internal platform data can be a valuable tool for improving downstream tasks like RTO.

4.2 Data and Social Re-Identification

In this section, we first provide the details of our ground truth RTO dataset, followed by social re-identification candidate extraction (§ 4.2.2) and validation (§ 4.2.3) steps.

4.2.1 Ground Truth Data

We can extract ground truth from all past orders and their subsequent outcomes of the e-commerce platform. Orders are subjected to multiple internal models during checkout, which can induce unintended biases in the data. To prevent this, 5% of all orders are randomly set aside, as the *control set*, where no intervention is applied. Further, we extracted the cash on delivery orders from the control set, because we observed that orders with cash on delivery are more prone to RTO. All our experiments and benchmarking are performed against this set. Our experiments are performed on 6 months (November 2021 - April 2022) of data. First 5 months of data is used for training, and the following 1 month is used as a test set.

We ensure that our study design does not breach the privacy terms and conditions of our platform, or of the social media platforms used. As an extra layer of prevention, experiments shown in this work are performed only on users who explicitly decided to make their name and city locations[1] public on the platform. After all filtration, our final dataset includes 6,881 orders placed by 2,121 unique users. Out of all, 2,201 (32%) orders were RTOed.

4.2.2 Potential Candidate Extraction

The initial step of user re-identification is to reduce the infinite search space of social profiles to a few candidate profiles for a given user. Querying social media platform's search engine using the *name* and *location* of a user has been shown to narrow the candidate pool effectively [47, 36]. For every unique user in our dataset, we create a search query of format <*user name*> <*city name*> and retrieve results from the social platform's search engines and a leading web search engine. Top 10 results of the query are used as candidate profiles.

We use a popular professional networking social media platform as a source of our social data; since, along with general information, such platforms have specific information that can reflect socio-economic indicators. Only data explicitly made public on the platform by the user is collected and used. Out of total 2,121 unique users, we found potential candidate profiles for 1,091 users.

[1] Used for social re-identification, see § 4.2.2

4.2.3 Social Re-Identification

Literature shows that different social profile attributes like name, location, network, and language features can be used to find a match from candidate profiles [108]. Considering the asymmetry between e-commerce and social media platforms, all these attributes are not available on both the platforms. However, we are in a unique position to access various locations a user has ordered from in the past. [36, 99, 126, 34] showed that matching various location information in a user's profiles with candidate profiles can find correct matches with a high probability.

We perform candidate filtration using two attributes viz. names and locations. Firstly, any candidate profiles whose names do not match the source user are rejected. In the second step, given a source user u, we extract from the orders history a set L_u, defined as $\{l_u^1, l_u^2, ..., l_u^n\}$ where l_u^i is the i^{th} city u placed a order at. For each potential candidate profile of u, a similar location set L_u^c is defined as $\{l_c^1, l_c^2, ..., l_c^m\}$ where c denotes a candidate profile and l_c^i is a city location mentioned in c's social profile.

The Match score of candidate profile c with u (α_u^c) is defined as the ratio of location in social profiles also present in the source user location set. While calculating the intersection between the set of city names fuzzy matching was used to account for slight variation in spellings and syntax of city names. E.g., Delhi vs. New Delhi, or Bangalore vs. Bengaluru.

$$\alpha_u^c = \frac{|L_u^c \cap L_u|}{|L_u^c|} \tag{4.1}$$

A candidate profile is considered a match if α_u^c is above a predefined threshold θ. A user can be classified into three categories based on the number of matches received. 'No match' for users where no candidate profiles had a score above θ, an 'Exact match' where exactly 1 candidate profile had the matching score above θ, and 'Multiple matches' in which case we found more than one candidate profile who had match score above the threshold. Table 4.1 shows the percentage of users in each of three categories for different values of θ. Users in the 'No match' category were removed from the modeling step. In case of 'Multiple match', final feature value is obtained by averaging over all the matches. Results shown in this paper are calculated using $\theta = 0.6$, results for varying values of θ were consistent and are omitted due to lack of space.

4.3 RTO Model

We discuss the features used by our proposed model, the types of modeling techniques we experimented with, and the evaluation metrics used.

4.3.1 Features

We broadly divide the features used into three categories; 1) past trends, 2) social profile quantitative, and 3) social profile abstractive. The first category is derived from internal data, and the other two are extracted from social profiles.

Past trends: These features are derived from historical data. Each sample includes the ratio of RTO vs. total orders over the last 3 months and 1 year for the user, products in order, seller, and product category. Apart from this, location is also a robust socio-economic indicator; therefore, we extract the same trends for pin code, street, and city mentioned in the delivery address. Further, we observed a correlation between the RTO rate and the order time (specifically the hour and weekday). Hence hour of the day, weekday, and respective past trends are added to the feature list.

Social profile quantitative: As we identify social profiles for a user, we extract if the user is a student, number of jobs, number of educational degrees, and number of friends and followers. The count of jobs/degrees may not always be a good indicator of someone's professional position since some people spend a long time in the same jobs, whereas others often switch jobs. Pertaining to that, we add two features counting the total years a user has spent working and in education.

Social profile abstractive: We have extracted social features related to the quantity of experience and education of users. Research has shown that institutions of education and programs studied can significantly impact career success [100]. Similarly, two people with the same years of job experience can have widely different buying propensities based on what roles they are pursuing at which organizations. We hypothesize features capturing user's education institutes and job roles can assist in RTO prediction. Recently, contextual language models pretrained on large volumes of data, have captured and exploited complex relations

Table 4.1: Results of social re-identification for varying values of matching threshold θ.

Match Threshold θ	Exact Match	Multiple Match	No Match
0.1	81.49	18.51	0.00
0.2	81.31	18.51	0.18
0.3	79.58	18.51	1.91
0.4	76.21	18.41	5.38
0.5	71.01	18.41	10.57
0.6	68.92	17.68	13.40
0.7	65.91	17.50	16.59
0.8	64.63	17.41	17.96
0.9	64.36	17.41	18.23
1.0	42.57	17.41	40.02

well for downstream tasks [26, 11]. Following this, we extract the latest education institute, and the course pursued by a given user and pass this textual information via a pretrained Sentence-BERT [97] model to generate 387 dimension vectors. A similar vector is also created for the Job organization and designation the user had while placing the order.

4.3.2 ML Modeling

Most of our data is tabular making tree-based ensemble methods like Random forest and XGBoost the default choice. Recently, attention-based architecture like Tabnet [3] has been proposed claiming to outperform traditional tree-based models. We present results on both types of models.

Figure 4.2 shows our training setup. In the tree-based models, 387 dimension vector obtained for job and education are decomposed to lower dimensions using UMAP [78] to prevent overfitting. The final dimension after decomposition is treated as a hyperparameter. Finally, decomposed vectors are added to the table of quantitative features as columns and fed into the model. When experimenting with deep learning-based models, tabular features are passed through Tabnet to generate a feature embedding. Generated embedding is concatenated with sentence-BERT embeddings (see § 4.3.1) and passed into a series of fully connected layers. All models are hyperparameter tuned using random search over; 4-fold cross-validation over the training data is used for parameter selection.

4.3.3 Evaluation

We use precision and recall to evaluate the performance of our models, but at a large scale, even very small improvements in model performance can lead to measurable revenue benefits. Additionally, traditional metrics may not always fit well in business discourse. Highlighting this, we define a metric named *Goodness* on which our models are evaluated.

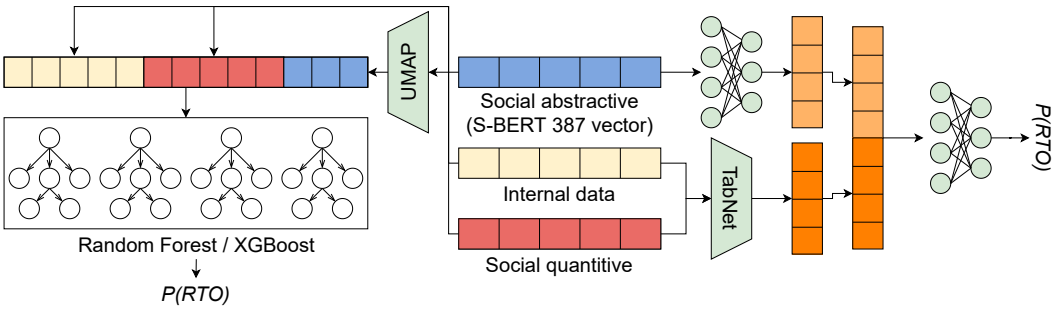


Figure 4.2: Our training architecture. In the case of tree-based models (on the left), all three feature sets are concatenated to form the input. While training deep learning models (on the right), tabular features are encoded via Tabnet and concatenated with S-BERT embeddings before being passed into a feed-forward neural network.

Goodness : It reflects the improvement in recall performance. Defined in Equation 4.2, it calculates the reduction in the ratio of RTO orders after being evaluated by the model. Multiplication with 10^4 is performed to convert value into Basis Points (bps), this improves readability even while observing quantitatively small improvements. A higher value is better.

$$Goodness = \left(\frac{|P|}{|P|+|N|} - \frac{|P| - |P_{Pred\ and\ True}|}{|P|+|N|-|P_{Pred}|} \right) \times 10^4 \tag{4.2}$$

$$FPR = \frac{|P_{Pred}| - |P_{Pred\ and\ True}|}{|P_{Pred}|} \tag{4.3}$$

Here, P is set of RTO orders, and N is set of Delivered orders. P_{Pred} is set of orders predicted as RTO by a model, and $P_{Pred\ and\ True}$ is set of true positive RTO predictions.

Our aim is to choose a classification threshold that maximizes *Goodness* while maintaining the false positive rate (FPR) below a fixed value. [2] A high FPR means increased false interventions, reducing customer experience. Just like precision and recall, *Goodness* and FPR are a trade-off balance. High *Goodness* comes with an increase in FPR.

4.4 Results

Table 4.2 shows performance of various RTO models on our test set. The random forest provides the overall best performance. As hypothesized, adding social features with past trends improves goodness by 300 bps, and adding contextual embeddings representing education and professional information improves the goodness further by 328 bps. This model has direct implications for improving the bottom-line revenue performance of an e-commerce organization.

Contrary to intuition, deep learning based models performed the worse. Comparative studies has shown that this behaviour is common in case of tabular data [37, 50, 109]. Studies compared the performance of Tabnet, and its contemporaries on a large variety of tabular data tasks, and concluded that these neural architectures do not perform consistently and are very sensitive to parameter tuning.

4.5 Conclusion and Future Work

Our study aims to improve the performance of a critical e-commerce problem RTO, where a user places an order and then cancels while the product is in transit, leading to logistics and opportunity cost. We hypothesize that fusing a users' social data with past RTO trend data can lead to improvements in performance. Towards this, we build a system to extract social profiles from popular professional networking social media platforms for a given user. Location-based matching is used to filter from the candidate matches. Finally, we extract quantitative and contextual features of matched profiles and demonstrate improvements of

[2] FPR threshold is decided based on product requirement.

3.1%, and 19.9% precision and recall, respectively, in the RTO detection task. Our work has direct implications for improving the bottom-line revenue of an e-commerce organization. Potential future directions of our work can be to experiment with transfer learning or multitask setup to see if social re-identification can help in other facets of e-commerce experience like review credibility or credit modeling. We would also like to extend our experiments to include data from a broader type of social media platforms.

Table 4.2: RTO detection performance on the test set. Random forest performs the best. The addition of social features with past trend data increases goodness by 628 bps.

Model	**Features**	**Precision (%)**	**Recall (%)**	**Goodness (bps)**
Random Forest	Past Trends	85.7	40.3	1,005.7
	Past Trends + Social quantitative	85.7	50.4	1,305.6
	Past Trends + Social quantitative + Social abstractive	**88.8**	**60.2**	**1,633.7**
XGBoost	Past Trends	80.0	33.6	809.3
	Past Trends + Social quantitative	82.2	39.7	994.1
	Past Trends + Social quantitative + Social abstractive	86.8	44.5	1,129.4
TabNet	Past Trends	82.4	39.4	977.0
	Past Trends + Social quantitative	78.2	30.2	716.2
	Past Trends + Social quantitative + Social abstractive	64.2	15.1	320.0

Chapter 5

Thesis Timeline and Outline

5.1 Timeline

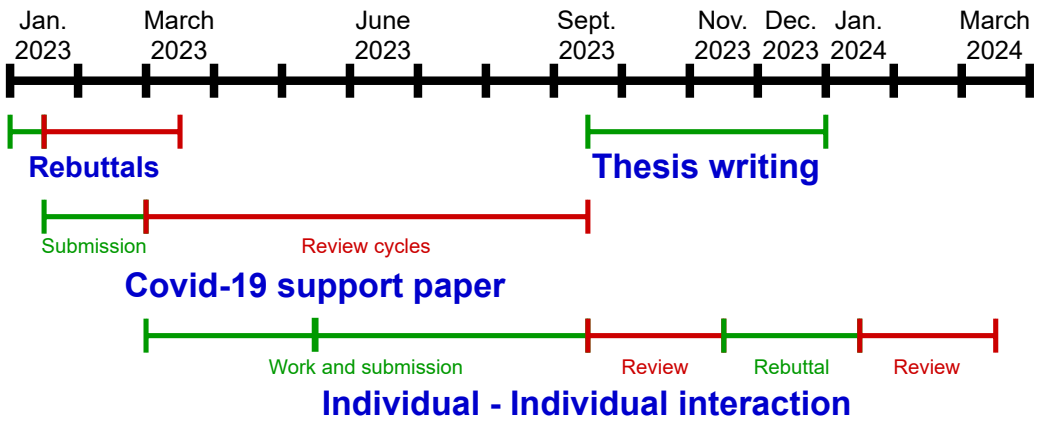


Figure 5.1: Potential thesis timeline.

5.2 Outline

Part I: Introduction and Background

1. Introductions
2. Preliminaries and Background

Part II: Individual - Community

3. Effect of Popularity Shocks on User Behavior
4. Effect of Feedback on Drug Consumption Disclosures
5. Effect of Social Support in Online Covid-19 Communities

Part III: Individual - Organization

6. Social Re-Identification Assisted RTO Detection for E-Commerce

Part IV: Individual - Individual

7. Study 5

Part V: Conclusion and Future Work

8. Conclusion

9. Future Work

Bibliography

[1] Nancy R Ahern, Penny Sauer, and Paige Thacker. Risky behaviors and social networking sites: how is youtube influencing our youth? *Journal of psychosocial nursing and mental health services*, 53(10):25–29, 2015.

[2] Tim Althoff, Pranav Jindal, and Jure Leskovec. Online actions with offline impact: How online social networks influence online and offline user behavior. In *WSDM*, 2017.

[3] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

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

[5] Nupur Baghel, Yaman Kumar, Paavini Nanda, Rajiv Ratn Shah, Debanjan Mahata, and Roger Zimmermann. Kiki kills: Identifying dangerous challenge videos from social media. *arXiv preprint arXiv:1812.00399*, 2018.

[6] Fabricio Balcazar, Bill L Hopkins, and Yolanda Suarez. A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, 7(3-4):65–89, 1985.

[7] Duilio Balsamo, Paolo Bajardi, and Andr?? Panisson. Firsthand opiates abuse on social media: monitoring geospatial patterns of interest through a digital cohort. In *The World Wide Web Conference*, 2019.

[8] Duilio Balsamo, Paolo Bajardi, Alberto Salomone, and Rossano Schifanella. 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*, 2021.

[9] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[10] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell,

Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[12] Christian Burgers, Allison Eden, Mélisande D van Engelenburg, and Sander Buningh. How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, 2015.

[13] Andrew T Carswell and Douglas C Bachtel. Mortgage fraud: A risk factor analysis of affected communities. *Crime, law and social change*, 52(4):347–364, 2009.

[14] Mehmet Cem Catalbas, Tomaz Cegovnik, Jaka Sodnik, and Arif Gulten. Driver fatigue detection based on saccadic eye movements. In *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 913–917. IEEE, 2017.

[15] Pei-Yu Chen, Samita Dhanasobhon, and Michael D Smith. All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. *Com (May 2008)*, 2008.

[16] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *ICWSM*, 2014.

[17] Thomas D Cook. "waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, pages 636–654, 2008.

[18] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972.

[19] Tiago Cunha, Ingmar Weber, and Gisele Pappa. 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*, 2017.

[20] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 99–103, 2016.

[21] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[22] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *WWW*, 2009.

[23] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the CHI 2016*, pages 2098–2110, 2016.

[24] Marijke De Veirman, Veroline Cauberghe, and Liselot Hudders. Marketing through instagram influencers: the impact of number of followers and product divergence on brand attitude. *International journal of advertising*, 36(5):798–828, 2017.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[27] Jay Dixit. Heartbreak and home runs: The power of first experiences, Jan 2010.

[28] Flavio Figueiredo. On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 741–746, 2013.

[29] Flavio Figueiredo, Jussara M Almeida, Fabrício Benevenuto, and Krishna P Gummadi. Does content determine information popularity in social media? a case study of youtube videos' content and their popularity. In *CHI*, 2014.

[30] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M Almeida. The tube over time: characterizing popularity growth of youtube videos. In *WSDM*, 2011.

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

[32] Karen Freberg, Kristin Graham, Karen McGaughey, and Laura A Freberg. Who are the social media influencers? a study of public perceptions of personality. *Public relations review*, 2011.

[33] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? In *ACM SIGMETRICS ICMMCS*, 2016.

[34] Xing Gao, Wenli Ji, Yongjun Li, Yao Deng, and Wei Dong. User identification with spatio-temporal awareness across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1831–1834, 2018.

[35] E Goffman. The presentation of self in. butler, bodies that matter. *The Presentation of Self in. Butler, Bodies that Matter*, 1959.

[36] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458, 2013.

[37] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[38] Amelia Grant-Alfieri, Judy Schaechter, and Steven E Lipshultz. Ingesting and aspirating dry cinnamon by children and adolescents: the "cinnamon challenge". *Pediatrics*, 131(5):833–835, 2013.

[39] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer, 2014.

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

[41] Catherine Hausman and David S Rapson. Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552, 2018.

[42] Atika Hermanda, Ujang Sumarwan, and Netti Tinaprillia. The effect of social media influencer on brand image, self-concept, and purchase intention. *Journal of Consumer Sciences*, 4(2):76–89, 2019.

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

[44] Han Hu, NhatHai Phan, James Geller, Stephen Iezzi, Huy T Vo, Dejing Dou, and Soon Ae Chun. An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. In *MedInfo*, 2019.

[45] Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

[46] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[47] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. @ i seek'fb. me' identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1259–1268, 2013.

[48] Soheil Jamshidi and Mahmoud Reza Hashemi. An efficient data enrichment scheme for fraud detection using social network analysis. In *6th International Symposium on Telecommunications (IST)*, pages 1082–1087. IEEE, 2012.

[49] S Venus Jin, Aziz Muqaddam, and Ehri Ryu. Instafamous and social media influencer marketing. *Marketing Intelligence & Planning*, 2019.

[50] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Regularization is all you need: Simple neural nets can excel on tabular data. *arXiv preprint arXiv:2106.11189*, 2021.

[51] Brian Keegan, Darren Gergle, and Noshir Contractor. Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–10, 2012.

[52] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: Structures and dynamics of wikipedia's coverage of breaking news events. *American behavioral scientist*, 2013.

[53] Brian C Keegan and Jed R Brubaker. 'is' to'was' coordination and commemoration in posthumous activity on wikipedia biographies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 533–546, 2015.

[54] Katherine Keith, David Jensen, and Brendan O'Connor. 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*, pages 5332–5344, Online, July 2020. Association for Computational Linguistics.

[55] Emre Kiciman, Scott Counts, and Melissa Gasser. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[56] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 EMNLP*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[57] Peter Kollock et al. The economies of online cooperation: Gifts and public goods in cyberspace. *Communities in cyberspace*, 239, 1999.

[58] Susanne Kopf. "rewarding good creators": Corporate social media discourse on monetization schemes for content creators. *Social Media+ Society*, 2020.

[59] Yusuf Kurniawan, Sri Kusumo Habsari, and Ismi Dwi Astuti Nurhaeni. 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, 2013.

[60] Hemank Lamba, Momin M Malik, and Jurgen Pfeffer. A tempest in a teacup? analyzing firestorms on twitter. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015.

[61] Hemank Lamba, Shashank Srikanth, Dheeraj Reddy Pailla, Shwetanshu Singh, Karandeep Singh Juneja, and Ponnurangam Kumaraguru. Driving the last mile: Characterizing and understanding distracted driving posts on social networks. In *Proceedings of the ICWSM*, volume 14, pages 393–404, 2020.

[62] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *European conference on information retrieval*, pages 689–695. Springer, 2016.

[63] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[64] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, 2014.

[65] Mark R Leary, Lydia R Tchividijian, and Brook E Kraxberger. Self-presentation can be hazardous to your health: impression management and health risk. *Health Psychology*, 13(6):461, 1994.

[66] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 2010.

[67] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. "we dissect stupidity and respond to it": Response videos and networked harassment on youtube. *American Behavioral Scientist*, 65(5):735–756, 2021.

[68] Xitong Li. How does online reputation affect social media endorsements and product sales? evidence from regression discontinuity design. In *WISE*, 2013.

[69] Xin Jean Lim, AM Radzol, J Cheah, and Mun W Wong. The impact of social media influencers on purchase intention and the mediation effect of customer attitude. *Asian Journal of Business Research*, 7(2):19–36, 2017.

[70] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[71] John Lu, Sumati Sridhar, Ritika Pandey, Mohammad Al Hasan, and Georege Mohler. Investigate transitions into drug addiction through text mining of reddit data. In *KDD*, 2019.

[72] Stephen Lyng. Edgework: A social psychological analysis of voluntary risk taking. *American Journal of Sociology*, 95(4):851–886, 1990.

[73] Danaja Maldeniya, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Herding a deluge of good samaritans: How github projects respond to increased attention. In *Web Conference*, 2020.

[74] Momin M Malik and Jürgen Pfeffer. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[75] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[76] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 197–201, 2016.

[77] Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

[78] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[79] David L McMillen and James B Austin. Effect of positive feedback on compliance following transgression. *Psychonomic Science*, 1971.

[80] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. " nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of international conference on multimedia retrieval*, pages 385–391, 2014.

[81] Ali Mert Ertugrul, Yu-Ru Lin, and Tugba Taskaya-Temizel. Castnet: Community-attentive spatio-temporal networks for opioid overdose forecasting. *arXiv e-prints*, 2019.

[82] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[83] R.G. Miller. *Survival Analysis*. Wiley Classics Library. Wiley, 2011.

[84] Ravita Mishra. Entity resolution in online multiple social networks (@ facebook and linkedin). In *Emerging Technologies in Data Mining and Information Security*, pages 221–237. Springer, 2019.

[85] Ahmet Anıl Müngen, Esra Gündoğan, and Mehmet Kaya. Identifying multiple social network accounts belonging to the same users. *Social Network Analysis and Mining*, 11(1):1–19, 2021.

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

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

[88] Vedant Nanda, Hemank Lamba, Divyansh Agarwal, Megha Arora, Niharika Sachdeva, and Ponnurangam Kumaraguru. Stop the killfies! using deep learning models to identify dangerous selfies. In *Companion Proceedings of the The Web Conference 2018*, pages 1341–1345, 2018.

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

[90] NIH. Overdose death rates. *National Institutes of Health*, 2021.

[91] Shrey Nougarahiya, Gaurav Shetty, and Dheeraj Mandloi. A review of e–commerce in india: The past, present, and the future. *Research Review International Journal of Multidisciplinary*, 6(03):12–22, 2021.

[92] Hüseyin Oktay, Brian J Taylor, and David D Jensen. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9, 2010.

[93] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009, 2015.

[94] Justin W Patchin and Sameer Hinduja. *Cyberbullying prevention and response: Expert perspectives*. Routledge, 2012.

[95] Cynthia M Pavett. Evaluation of the impact of feedback on performance and motivation. *Human Relations*, 36(7):641–654, 1983.

[96] Jonatas Pulz, Renan B Muller, Fabio Romero, André Meffe, Álvaro F Garcez Neto, and Aldo S Jesus. Fraud detection in low-voltage electricity consumers using socio-economic indicators and billing profile in smart grids. *CIRED-Open Access Proceedings Journal*, 2017(1):2300–2303, 2017.

[97] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[98] Howard Rheingold. *The virtual community: Finding commection in a computerized world*. Addison-Wesley Longman Publishing Co., Inc., 1993.

[99] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. Linking users across domains with location data: Theory and validation. In *Proceedings of the 25th international conference on world wide web*, pages 707–719, 2016.

[100] Georgeanna FWB Robinson, Lisa S Schwartz, Linda A DiMeglio, Jasjit S Ahluwalia, and Janice L Gabrilove. Understanding career success and its contributing factors for clinical and translational investigators. *Academic medicine: journal of the Association of American Medical Colleges*, 91(4):570, 2016.

[101] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

[102] Lauren O Roussel and Derek E Bell. Tweens feel the burn:“salt and ice challenge” burns. *International journal of adolescent medicine and health*, 28(2):217–219, 2016.

[103] J Philippe Rushton and Goody Teachman. The effects of positive reinforcement, attributions, and punishment on model induced altruism in children. *Personality and Social Psychology Bulletin*, 1978.

[104] Jorma Saari and Merja Näsänen. The effect of positive feedback on industrial housekeeping and accidents; a long-term study at a shipyard. *International Journal of Industrial Ergonomics*, 4(3):201–211, 1989.

[105] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 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, pages 440–451, 2019.

[106] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[107] Hanan Shteingart, Tal Neiman, and Yonatan Loewenstein. The role of first impression in operant learning. *Journal of Experimental Psychology: General*, 142(2):476, 2013.

[108] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter*, 18(2):5–17, 2017.

[109] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[110] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. Was this review helpful to you? it depends! context and voting patterns in online content. In *WWW*, 2014.

[111] BF Skinner. The behavior of animals: An experimental analysis. *New York: Appleton-Century Crofts*, 1938.

[112] Burrhus Frederic Skinner. The behavior of organisms: an experimental analysis. In *Appleton-Century*, 1938.

[113] Crystal R Smit, Laura Buijs, Thabo J van Woudenberg, Kirsten E Bevelander, and Moniek Buijzen. The impact of social media influencers on children's dietary behaviors. *Frontiers in psychology*, 10:2975, 2020.

[114] Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*, 2019.

[115] Elizabeth A Stuart. 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, 2010.

[116] Acar Tamersoy, Duen Horng Chau, and Munmun De Choudhury. Analysis of smoking and drinking relapse in an online community. In *Proceedings of the 2017 international conference on digital health*, pages 33–42, 2017.

[117] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[118] Sharon Tennyson. Economic institutions and individual ethics: A study of consumer attitudes toward insurance fraud. *Journal of Economic Behavior & Organization*, 32(2):247–265, 1997.

[119] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 1960.

[120] Edward L Thorndike. Animal intelligence. *Nature*, 58(1504):390–390, 1898.

[121] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[122] Ebru Uzunoğlu and Sema Misci Kip. Brand communication through digital influencers: Leveraging blogger engagement. *International journal of information management*, 2014.

[123] Marat Valiev, Bogdan Vasilescu, and James Herbsleb. Ecosystem-level determinants of sustained activity in open-source projects: A case study of the pypi ecosystem. In *ESEC/FSE*, 2018.

[124] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[125] I Veissi. *Influencer Marketing on Instagram. Yayınlanmış Lisans Tezi, Haaga-Helia University Of Applied Sciences*. PhD thesis, Bachelor's Thesis Degree Programme in International business, 2017.

[126] Huandong Wang, Yong Li, Gang Wang, and Depeng Jin. Linking multiple user identities of multiple services from massive mobility traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(4):1–28, 2021.

[127] Ke Wang, Mohit Bansal, and Jan-Michael Frahm. Retweet wars: Tweet popularity prediction via dynamic multimodal regression. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1842–1851. IEEE, 2018.

[128] Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference, 2021.

[129] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 2013.

[130] Reza Zafarani and Huan Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49, 2013.

[131] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel Robert, and Daniel Romero. Shocking the crowd: The effect of censorship shocks on chinese wikipedia. In *ICWSM*, 2017.

[132] Ark Fangzhou Zhang, Ruihan Wang, Eric Blohm, Ceren Budak, Lionel P Robert Jr, and Daniel M Romero. Participation of new editors after times of shock on wikipedia. In *ICWSM*, 2019.

[133] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1513–1522, 2015.