Down the Rabbit Hole: Modeling Twitter Dynamics through Bayesian Inference

Zachary Novack*
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
znovack@andrew.cmu.edu

Simon DeDeo
Department of Social and Decision Science
Carnegie Mellon University
Pittsburgh, PA 15213
sdedeo@andrew.cmu.edu

Abstract

Social media usage, and its impact on people's physical and mental health, is of interest to a diverse range of academic disciplines and everyday people. Despite this, we know very little about the ways in which, over months and years, someone's social media use may escalate to consume significant amounts of their daily leisure time. Nor do we understand the ways in which a user's posts may shift into toxic or unexpectedly abusive patterns. Understanding the long-term dynamics of use is complicated by the fact that day-to-day engagement has significantly non-normal statistics and may fluctuate by orders of magnitude—informally, users are sometimes driven to rare "binges" with lasting consequences for their future trajectory. To address this complex interplay of timescales, this work presents a Bayesian model for usage over time, flexible enough to capture a wide range of short and long-term temporal dependencies. Examining the "dose response" curves of a random sample of 500 users, we find that most users ($\approx 90\%$) show evidence for a stable, equilibrium level of use. A smaller "high-risk" subset ($\approx 10\%$) show evidence for instability: when short-term fluctuations drive their levels of use sufficiently high, they enter a new phase of sustained, run-away usage. Once we control for the levels of use, we find that "likes", retweets, and other forms of feedback received from other users do not significantly impact future behavior. This casts doubt on the common heuristic that social media use is driven by an "addiction to likes": for example, there is no evidence that a user whose posts receive an unexpectedly low level of likes posts more to "make up the difference". Finally, in looking at the dynamics of user toxicity, we find a tail-risk effect: prior toxic behavior rarely shifts the user's median post, but rather increases the likelihood of (otherwise rare) extreme toxicity. Our flexible dynamical modeling approach reveals significant heterogeneity in the ways in which users adapt to social media systems, and opens the door for more qualitative investigations into the outsized effects social media may have on users across arbitrarily long timescales.

1 Introduction

Social media use is ubiquitous in modern society [3]. While many users derive great benefits from the ability to connect with others online, there has been increasing attention to the negative consequences of engagement, including the outsized mental and emotional effect on users that come from heavy use. Negative effects include maladaptive repercussions such as depression, anxiety [11, 20], worsened work performance [29], and political polarization [2, 15].

Maladaptive social media usage, often referred to as social media addiction (SMA, though this term's legitimacy has been raised given its connection to physical addiction), has been approached through a

*zacharynovack.github.io

variety of different scientific paradigms [24]. This has produced a sea of different models of SMA, with very little attempt at reconciling these contrasting approaches into a holistic definition of SMA and unsafe social media use.

Before we can even begin to formally diagnose how SMA may arise, however, we still lack a clear understanding of the dynamics of social media usage over time. While there is a heuristic sense that "people love to use social media," how such usage evolves over time has been given very little attention in the academic sphere. We still know little about long-term usage patterns (*e.g.*, whether users exhibit stable equilibria, complete runaway effects, or some combination of the two) and the temporal patterns for *other* salient aspects of social media (such as hate speech and toxicity). Additionally, much work in SMA research has highlighted the importance of perceived engagement on usage [26, 23, 18], but such claims lack empirical validation and analysis through social media usage in the wild.

One of the difficulties with SMA research is the existence of extreme and "bursty" fluctuations in usage over time: it is not uncommon, for example, for a user who shows a steady level of a few posts a day to suddenly engage at a much higher rate, many standard deviations above average—and just as suddenly to return to lower levels of use. This complicates analyses that rely on the assumption of normally-distributed statistics. A second difficulty is that these sudden bursts of activity might well influence the user's behavior on longer timescales but, again, standard statistical tools have difficulty modeling how a "binge" might be followed by a "purge" (*i.e.*, drastically lowered levels of use) or, conversely, might more permanently elevate baseline usage.

To address this gap, we propose a novel methodology for analyzing social media usage time series that is readily interpretable by domain experts, and is able to capture a surprisingly expressive range of possible usage patterns.

The results of our modeling process highlight a few main takeaways:

1. We find that we may broadly categorize users into low, medium, and high risk for unhealthy behavior groups, based on the dynamics of our estimated parameters for each user, and that most users on Twitter fall within the low-to-no risk category.
2. We find that perceived engagement has no significant separable effect on either user posting frequency or toxicity, even when accounting for adjustment to baselines. While this doesn't disprove possible attention mechanisms as a driving factor, it calls into question previous conclusions of reaction to engagement effects on social media.
3. We also present an initial exploration of toxicity dynamics on Twitter, and find that all investigated users showed very low toxicity rates, and that dynamical variations only seem to fatten the tailed-ness of the toxicity distribution to moderate levels.

2 Previous Work

The general landscape for work on social media usage and its external effects is broad and incredibly diverse. Namely, there exists a large psychology-driven wing that seeks to explain maladaptive social media use through traditional psychological avenues. Work has been done to analyze social media use in controlled experimentalal settings through attachment theory [26, 6], gratification theory [9], planned behavior theory [13], and stimulus-response theory [27].

Outside of the traditional psychological theory, there has been work looking at social media use from neurological perspectives [23], economic perspectives [1], and in placing the role of technology and social media directly at the core of new addiction models [25, 19, 28]. These all offer interesting possible addiction models and some noticeable empirical results in the randomized control setting, though there is little common thread between most of these approaches. Additionally, the lack of exploration of social media in *non-control* settings, and in true longitudinal cases, presents a clear gap in all of these papers.

There has also been a vein of research attempting to apply machine learning (ML) approaches to social media use, which is more in line with our methodological approach. There is an emerging trend of ML being used for drug-related addiction research [17, 7], though work on specifically social media addiction has seen comparably less attention. [21] and [18] both provide limited experimental studies in which modern ML techniques were used to predict maladaptive social media use.

These endeavors provide a decent methodological foundation, though the adherence to traditional supervised learning approaches adds considerable assumptions to the modeling process (this problem is true also in ML approaches for traditional addictive substances, as noted by [17]). Our method, though it is "supervised" in the sense that we learn our model using matched pairs of predictor and response variables, offers a departure from these approaches by assuming *nothing* about the underlying definition or processes of social media addiction, and focuses only on truly observable metrics (in this, being "unsupervised" with regards to modeling addiction).

With regards to measuring toxicity (and its wider implications) on social media, there has been a large body of work on estimating text toxicity within the ML domain [10, 4], and additionally in measuring polarization dynamics on social media, which are intrinsically linked to toxicity trends [15, 2]. These works provide us with a robust methodological baseline to perform our investigation of toxicity dynamics for individual users across time scales, which to our knowledge has never been investigated before using state-of-the-art estimation approaches.

3 Data and Methods

3.1 Data Collection

Our research focuses specifically in modeling the dynamics of usage on the social networking site **Twitter**. Twitter sports an easy-to-use public API, and given its relatively friendly data structure (being a mostly text-based medium), it is often the website of choice for most hands-on social media research [5]. By leveraging the academic tier of the public API, we are able to attain the *entire* public posting history for a given user. We produced a pseudo-random sample of 500 users by randomly sampling from users who had recently interacted with the top 20 most followed Twitter accounts (e.g. Barack Obama, Ariana Grande, CNN, etc.), which became the basis for the empirical results of the present paper.

3.2 Data Structure and Preprocessing

For every user in our dataset, we collected the full time-series of public posts they ever made. On Twitter, there are four basic kinds of posts:

- *Tweets:* The main form of Twitter usage, comprising of at most 280 unicode characters and optional media attached (i.e. images or video)
- *Replies:* Comments that are made to an original post's thread
- *Retweets:* Shared versions of other users' posts that contain no additional text from the sharer
- *Quotes:* Retweets with added text by the sharer

In order to understand the mechanisms of usage on social media *with regards to the content directly generated by each user*, we chose to remove retweets from our analyses. Thus, for every user, we produce a time series of total *daily* posts made (excluding retweets). We center the time frame of this, and all subsequent time series such that the hour with the lowest average posting per day becomes midnight.

Additionally, there are also four basic types of engagement metrics that Twitter keeps track of for each post: likes, replies, retweets, and quotes. We group these metrics into two broad categories: **Positive Valence Responses** or **PVR** (likes and retweets, as such responses contain no text data and thus generally only confer positive engagement), and **Mixed Valence Responses** or **MVR** (replies and quotes, as the valence of such responses is unknown without attempting to directly estimate the sentiment). For the purposes of the present study, we specifically focused on the **Average PVR (APVR)** within a given time frame, as such a metric is more robust to changes in posting frequency and avoids possibly erroneous sentiment estimation. For every user, we z-score their APVR time-series using a month-long backwards-facing sliding window in order to encapsulate the possible habituation effect [22, 16], in which users' need for the addictive medium is generally mediated by recent previous usage levels.

On top of the posting frequency time series and APVR time series, we too look to estimate how user toxicity evolves as a result of past behavior and perceived engagement. Using a pretrained

RoBERTa-based model [12] that estimates text toxicity (while balancing for unintended bias in such estimations), we thus generate a time-series of average daily toxicity scores for each user, where toxicity is constrained between 0 and 1 inclusive (as it is a measure of the "probability" that a given post is toxic).

3.3 Methods

3.3.1 Bayesian Integer Autoregression

The basis of our quantitative approach is based in Bayesian inference through the programming language STAN [8]. The advantages of a Bayesian approach to this problem are manifold. Consider first, by contrast, a standard integer-based *frequentist* degree-1 autoregression: Given a sequence of values $Y_1, Y_2, \ldots, Y_T$, we attempt to learn a set of parameters $\boldsymbol{\beta}$ to some linear function $f(x; \boldsymbol{\beta})$ such that:

$$Y_t = \exp(f(Y_{t-1}; \boldsymbol{\beta})) \tag{1}$$

In this sense, we implicitly assume that each Y_t follows a Poisson distribution with mean $f(Y_{t-1}; \boldsymbol{\beta})$. The degree here notes the number of previous time step terms we may include in our parameterization. There are a few things to note in such an approach. Namely, such frequentist estimation obfuscates the stochastic process of such a time-series, and a priori assumes that the *parameter* distributions are of some standard form (typically Gaussian). Additionally, the implicit Poisson assumption additionally assumes no overdispersion, or more formally that the variance of the distribution is reasonably identical to the mean, which is large assumption in practice.

In a *Bayesian* approach, we are now focused with learning the *distribution* over our predicted time-series, and we additionally directly specify the prior distribution over our latent variables $\boldsymbol{\beta}$. In order to combat the possible overdispersion behavior in our time-series, we can assume that our Y_t's come from a *Negative Binomial* distribution. While the Negative Binomial distribution is classically known to model the number of successes (each with probability p) before r failures, we can alternatively parameterize it with its mean μ and dispersion parameter ϕ, where if $Y \sim \text{NegBin}(\mu, \phi)$ then:

$$P(Y = y) = \frac{\Gamma(\phi + y)}{y!\Gamma(\phi)} \left(\frac{\phi}{\phi + \mu}\right)^{\phi} \left(\frac{\mu}{\phi + \mu}\right)^{y} \tag{2}$$

For some positive integer y and where $\Gamma(x)$ is the gamma function. Note that as $\phi \to \infty$, $\text{NegBin}(\mu, \phi) \to \text{Pois}(\mu)$, and thus Negative Binomial regression acts as a strictly more expressive form of integer-based modeling than Poisson regression. Thus, in a typical Bayesian model specification to mirror (1), we would have that:

$$\begin{aligned} \boldsymbol{\beta} &\sim P(\boldsymbol{\theta}) \\ \phi &\sim Q(\boldsymbol{\psi}) \\ Y_t &\sim \text{NegBin}(f(Y_{t-1}; \boldsymbol{\beta}), \phi) \end{aligned} \tag{3}$$

For some prior distributions P and Q parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. Notice that by directly specifying the prior distributions on our latent variables and focusing on learning distributions rather than point estimates, we are able to attain much more reliable uncertainty estimates on our parameters and model fit in general.

3.3.2 Activity Model

We can now directly present our given model for user activity. For a given user with time series of daily number of posts $\{Y_t\}_{t=1}^{T}$ (excluding retweets) and matched APVR time series $\{X_t\}_{t=1}^{T}$, we say that:

$$\begin{aligned} \beta_0, \beta_1, \beta_2, \gamma &\sim \text{Laplace}(0, 1) \\ \phi &\sim \text{Exp}(1) \\ Y_t &\sim \text{NegBin}(\text{ReLU}(\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-1}^2 + \gamma X_{t-1}), \phi) \end{aligned} \tag{4}$$

Table 1: Estimated γ Values for Activity Model

γ	Attention Effect	% of Dataset
$\gamma \approx 0$	No Response	66.5
$\gamma > 0$	Increases w/Increasing Attention	21.1
$\gamma < 0$	Increases w/Decreasing Attention	12.4

Where $\text{ReLU}(x) = \max(0, x)$ is used to allow for possibly negative latent parameters while enforcing the nonnegativity constraint on the mean of the Negative Binomial distribution. Note that parameterizations with added time series are normally referred to as *exogenous* autoregressive models. This specific parameterization performed better or as good as a suite of other similar model specifications tested (including fitting interaction terms, exponential coefficients, and/or hardline maximums on the mean of the distribution) in terms of log-likelihood fit. Additionally, the presence of the quadratic term allows for an intuitive understanding of user behavior through traditional psychological means, as shown in the next section.

3.3.3 Toxicity Model

For modeling user toxicity, our time series of interest is now constrained within the range of real numbers $[0, 1]$, and not positive integers like before. Similar to how we can parameterize the Negative Binomial distribution in terms of μ and ϕ, we can parameterize the *Beta* distribution (which being defined on $[0, 1]$ makes it an intuitive candidate distribution for estimating the distribution of probabilities) in terms of mean μ and sample size ν, such that if $T \sim \text{Beta}(\mu, \nu)$ them:

$$P(T = x) = x^{\mu\nu-1}(1-x)^{\nu-\mu\nu-1}\left(\frac{\Gamma(\nu)}{\Gamma(\mu\nu)\Gamma(\nu-\mu\nu)}\right) \tag{5}$$

In this parameterization, ν has the nice property of qualitatively controlling the "shape" of the distribution, with larger ν values leading to a clear centroid around μ while smaller ν values dictation a right-skewed distribution.

Thus, for a given user and toxicity time series $\{T_t\}_{t=1}^{T}$ and matched APVR time series $\{X_t\}_{t=1}^{T}$, we say that:

$$\begin{aligned} \alpha_0, \alpha_1, \alpha_2, \gamma &\sim \text{Laplace}(0, 1) \\ \nu &\sim \text{Exp}(10) \\ T_t &\sim \text{Beta}(\text{ReLU}(\alpha_0 + \alpha_1 T_{t-1} + \alpha_2 T_{t-1}^2 + \gamma X_{t-1}), \nu) \end{aligned} \tag{6}$$

Where once again the ReLU term is used to enforce nonnegativity in the distribution mean.

For each user, we learn the parameters $\beta_0, \beta_1, \beta_2, \gamma, \phi$ for the Activity model and $\alpha_0, \alpha_1, \alpha_2, \gamma, \nu$ for the toxicity model using the No-U-Turn-Sampler [14], or NUTS method, a variant of Hamiltonian Monte Carlo that eliminates the need for prior specification on the number of steps to evolve the Hamiltonian at each main time step. We run the NUTS for each model with 4 chains of 1000 iterations each using the PyStan API.

4 Results

4.1 Perceived Engagement Dynamics

Of key importance of this modeling procedure was to observe the effect that perceived positive engagement, the X_t series, had on posting dynamics. As we can see in Table 1, about two thirds of users in our dataset had estimated γ parameters that were not significantly different from 0. In the small minority of users who did have significant nonzero estimated γ values, we found that the average absolute difference between the estimated negative binomial mean with and without the γX_{t-1} term was heavily skewed towards zero, and never got above ≈ 4.

This highlights one of our major findings: **we find little evidence of any significant separable effect that perceived positive engagement has on posting dynamics**. This calls into doubt previous theories of SMA that propose modeling SMA through "engagement" as an addictive medium, and thus may offer a somewhat more positive view of the social media landscape. Hence, for the rest of the paper we disregard the γX_{t-1} and focus soley on the autoregressive mechanisms of previous activity at play.

4.2 Activity Dynamics and Equilibria

We begin with a case study of four users in our dataset, labeled users i, j, k, and ℓ respectively for sake of anonymity. In using a quadratic parameterization for our model, we can analyze the dose-response curve for any given user in a deterministic setting (that is, if $Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-1}^2$ explicitly).

In this setting, we find multiple different equilibrium behaviors based on when, and how many times the dose-response curve intersects with the $Y_t = Y_{t-1}$ line (which would yield a completely stationary pattern at any value of Y_{t-1}). We broadly characterize these behaviors in the following definition:

Definition 1 *For any dose-response function $f(Y_t)$ and point Y_t' such that $f(Y_t') = Y_t'$, if $\left|\frac{df}{dY_t}|_{Y_t=Y_t'}\right| < 1$, then Y_t' is an **attractor**. If $\frac{df}{dY_t}|_{Y_t=Y_t'} > 1$ then Y_t' is an **repeller**. If $\frac{df}{dY_t}|_{Y_t=Y_t'} < -1$, then Y_t' is a **chaotic equilibrium**.*

For attractors, iterating the function will eventually converge to the Y_t' value as long as there isn't another equilibrium point between the initial value and Y_t' (for dose-response curves with only one equilibrium point, starting anywhere will converge to the attractor. In the positive-slope repeller case, iterating the function will yield points further and further away from the repeller given the starting point, until the function either hits another equilibrium point or explodes completely. Chaotic equilibria are an interesting theoretical case, as changes in the slope at such an equilibria can yield both cyclic patterns and purely chaotic behavior as dictated by the appropriately scaled logistic map.

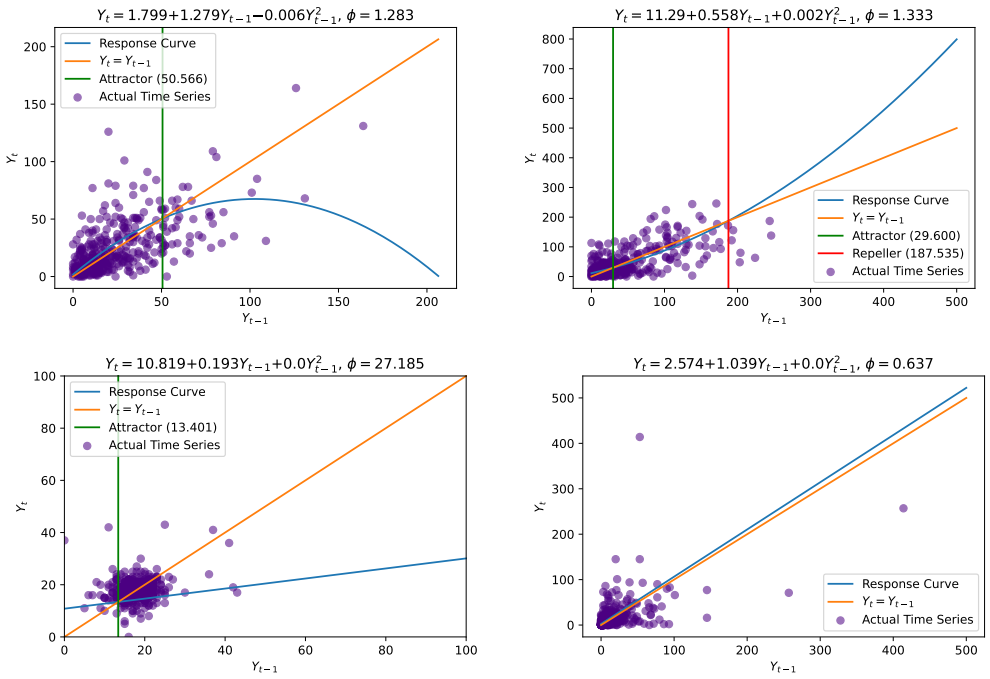
However, what we actually care about are the dynamics of the system in our modeled *stochastic* setting, where each point sets the mean of a corresponding negative binomial distribution to be sampled from. In this regime, we have the extra dispersion parameter ϕ, which implicitly controls the shape of the distribution. When ϕ is small (especially when $\phi \in [0, 1]$), the distribution is severely right-skewed and shows no clear centroid around the mean. As ϕ increases, the distribution develops a centroid around its mean μ ,which gradually flattens as $\mu \gg \phi$.

In Figure 4.2, we plot the theoretical response curve along with the $Y_t = Y_{t-1}$ line and note any equilibrium points, and additionally plot the last year's worth of post data (in purple). Users i and k are similar in their status as single-attractors, though they differ in terms of the ϕ values (as k's is much larger than i's) and in the shape of the response curve (k doesn't have any estimated effect for the Y_{t-1}^2 term). We see these differences play out rather clearly in the actual time series, as the attractor for user i is more of an "upper limit" of common use than an centroid (which is what we'd expect with a low ϕ value), and random extreme points are mostly met with immediate collapse back to low values (which we would expect with the quadratic response curve). For user k, the simulated series is clearly centered about the attractor, and more extreme points are met with similar values as they converge back to the equilibrium. Given this, we could reasonably characterize users i and k as having "safe", self-moderating levels of usage.

Users j and ℓ are a different story. In user j, which has both an attractor *and* a repeller equilibrium, most of the simulate points are clustered towards low values (as with the low ϕ any clear centering about the attractor isn't immediately clear), though posts further away from the attractor (and closer to the repeller) happen at decent frequently and are rather close to the $Y_t = Y_{t-1}$ line (indicating only slow regression to the attractor). User ℓ is even worse off, as with no equilibrium point the user has multiple posting days with incredibly high values. In this sense, we could characterize j and ℓ as being at risk for *unhealthy* usage.

Moving beyond these four examples, we can build a qualitative taxonomy of users by grouping them into the following categories:

- **Fixed Repellers (FR)**: Users with *no* positive equilibrium point, and thus show runaway behavior that is *only* moderated by stochastic shocks to the system



b

Figure 1: Dose-Response Curves (blue), with vertical lines for equilibria points (green and red) and actual time series (purple) for users i (top left), j (top right), k (bottom left), and ℓ (bottom right).

Table 2: Taxonomy of users by dynamic properties

Group	% of Dataset
Stable Attractors	89.45
Bi-Modal Attractor-Repellers	7.30
Fixed Repellers	3.25

- **Stable Attractors (SA)**: Users with *one* attractor point, where usage is moderated not only by stochastic shocks but also internal moderation effects to stay within some reasonable range of posting frequency
  - This group could be further split by response curve type (quadratic vs. linear, wherein the former should exhibit more boom-bust mechanics while the latter shows more gradual regression to stability) and by ϕ value
  - Note that we also group users who have two equilibria with the higher of the two being an *attractor* into this category ($n = 1$), as such behavior is mostly identical to that of the 1 equilibrium attractor case.
- **Bi-Modal Attractor Repellers (BMAR)**: Users with *two* equilibria, a lower attractor and higher repeller, where usage is internally moderated below the repeller, but may exhibit runaway chaotic behavior past the repeller cutoff.

In Table 4.2, we can see that users in the Stable Attractor group make up nearly 90% of our dataset, withe 7% in the bi-modal group and just 3% in the fixed repeller group (note that though our model is expressive enough to capture cyclic and chaotic behavior, we did not find any users in our dataset that showed this behavior and thus excluded such behavior in our taxonomy). At a high level, this result highlights that **while most users on Twitter may exhibit stable, "healthy" behavior, there is a nonzero population of users who are at risk for possibly unhealthy usage**.

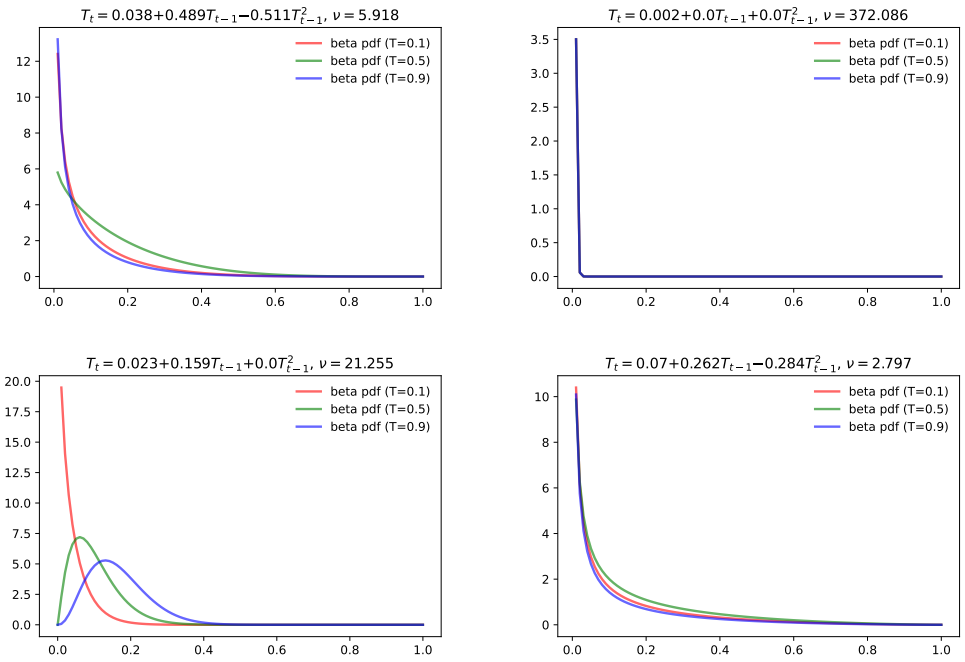


Figure 2: Estimated Beta distributions at $T = \{0.1, 0.5, 0.9\}$ for users i (top left), j (top right), k (bottom left), and ℓ (bottom right).

4.3 Toxicity Dynamics

Given compute restrictions and the slow inference time of our chosen Toxicity model, we focus our results here on our small case study of users i, j, k and ℓ once more. Note additionally for these users that the γX_{t-1} terms are all insignificant (as we saw before with the Activity model), and thus we disregard this term once again. While we can construct similar dose-response simulations (as in Figure 4.2) for the toxicity model, we may also directly visualize the resultant Beta distributions at different prior toxicity values. Figure 4.3 shows exactly this for our case study of users.

Namely, the main common thread in this limited case study is the fact that **no amount of prior toxicity shifts the mass of the distribution to the toxic side of the spectrum**. Within this common thread, we do see a wide range of behaviors: the stationary distribution of user j is one thing to note, wherein the estimate ν is high enough to predict mass nearly *only* at the mean. Users i, k, and ℓ show more dynamic behavior, with users i and ℓ showing quadratic response behavior with low ν values (and thus a skewed heavy-tailed distribution) and user k showing linear response behavior with a higher ν value (allowing for a centroid to appear around the mean).

It is interesting to note that in comparing both the Toxicity and Activity models, our users with unstable activity patterns (j and ℓ) show completely relatively stable toxicity patterns, while our users with more stable activity (i and k) display much more dynamic toxicity behavior. This possibly raises the hypothesis that users at risk of unhealthy posting patterns may not be shifting their posting *content* in any maladaptive way, and that users with more healthy posting patterns may still show spikes of maladaptive behavior through the content of their posts, though more research is needed to confirm this on a wider sample of users.

5 Discussion

At a high-level, our work presents a robust modeling technique to the realm of social media addiction research, which is both simple and interpretable, and is expressive enough to capture an incredibly diverse range of potential user behavior patterns that extend to long time scales. In using this method, we find that the state of social media use may not be as bleak as some make it out to be: most users

do show healthy levels of use, and don't seem to react to perceived engagement in any meaningfully separable way. This being said, we are also able to identify a nonzero population of users who are at risk of potentially unhealthy usage on Twitter.

There are multiple avenues of further work that this project may motivate. Namely, while our initial foray into modeling toxicity dynamics is insightful, more work must be done to produce generalizable conclusions on common patterns between the toxicity of users (much like has been done in the activity model), and to link to the two modeling approaches to create a more robust taxonomy of Twitter usage as a whole. Additionally, there may be interested in extending the current parameterization of our models to higher order polynomial terms, as such would allow for more complex equilibria behavior that may capture the dynamics of usage or toxicity better than our quadratic model. Approximation of the mean of the predictive distribution could even be done through more sophisticated approaches utilizing recurrent neural networks (RNNs), though such explorations may trade off expressivity for interpretability.

Given our modeling approach, we hope this opens the door for stakeholders, such as psychologists and policy makers, to utilize our model for more qualitative analyses of maladaptive social media use as whole. Since our model assumes *nothing* about the true underlying mechanisms of addiction and only provides a way to predict usage dynamics, it is flexible enough to be used with a host of prior addiction definitions and down-stream priorities. Thus, the approach presented here may have profound implications for uniting the diverse field of researchers who care about social media use under a similar methodological paradigm.

References

[1] Hunt Allcott, Matthew Gentzkow, and Lena Song. *Digital Addiction*. Working Paper 28936. National Bureau of Economic Research, June 2021. DOI: 10.3386/w28936. URL: http://www.nber.org/papers/w28936.

[2] Hunt Allcott et al. "The welfare effects of social media". In: *American Economic Review* 110.3 (2020), pp. 629–76.

[3] Monica Anderson, Jingjing Jiang, et al. "Teens, social media & technology 2018". In: *Pew Research Center* 31.2018 (2018), pp. 1673–1689.

[4] Darko Andročec. "Machine learning methods for toxic comment classification: a systematic review". In: *Acta Universitatis Sapientiae, Informatica* 12.2 (2020), pp. 205–216.

[5] Despoina Antonakaki, Paraskevi Fragopoulou, and Sotiris Ioannidis. "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks". In: *Expert Systems with Applications* 164 (2021), p. 114006.

[6] L Badenes-Ribera et al. "Parent and peer attachment as predictors of Facebook addiction symptoms in different developmental stages (early adolescents and adolescents)". In: *Addictive behaviors* 95 (2019), pp. 226–232.

[7] Elan Barenholtz, Nicole D Fitzgerald, and William Edward Hahn. "Machine-learning approaches to substance-abuse research: Emerging trends and their implications". In: *Current opinion in psychiatry* 33.4 (2020), pp. 334–342.

[8] Bob Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017).

[9] Silvia Casale and Giulia Fioravanti. "Why narcissists are at risk for developing Facebook addiction: The need to be admired and the need to belong". In: *Addictive behaviors* 76 (2018), pp. 312–318.

[10] Navoneel Chakrabarty. "A machine learning approach to comment toxicity classification". In: *Computational intelligence in pattern recognition*. Springer, 2020, pp. 183–193.

[11] Rachel A Elphinston and Patricia Noller. "Time to face it! Facebook intrusion and the implications for romantic jealousy and relationship satisfaction". In: *Cyberpsychology, behavior, and social networking* 14.11 (2011), pp. 631–635.

[12] Laura Hanu and Unitary team. *Detoxify*. Github. https://github.com/unitaryai/detoxify. 2020.

[13] Shirley S Ho, May O Lwin, and Edmund WJ Lee. "Till logout do us part? Comparison of factors predicting excessive social network sites use and addiction between Singaporean adolescents and adults". In: *Computers in Human Behavior* 75 (2017), pp. 632–642.

[14] Matthew D Hoffman, Andrew Gelman, et al. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[15] Ferenc Huszár et al. "Algorithmic amplification of politics on Twitter". In: *Proceedings of the National Academy of Sciences* 119.1 (2022).

[16] David Lloyd, Kathryn Hausknecht, and Jerry Richards. "Nicotine and Methamphetamine Disrupt Habituation of Sensory Reinforcer Effectiveness in Male Rats". In: *Experimental and clinical psychopharmacology* 22 (Apr. 2014), pp. 166–75. DOI: 10.1037/a0034741.

[17] Kwok Kei Mak, Kounseok Lee, and Cheolyong Park. "Applications of machine learning in addiction studies: A systematic review". In: *Psychiatry research* 275 (2019), pp. 53–60.

[18] Davide Marengo et al. "Mining Digital Traces of Facebook Activity for the Prediction of Individual Differences in Tendencies Toward Social Networks Use Disorder: A Machine Learning Approach". In: *Frontiers in psychology* 13 (2022), p. 830120. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.830120. URL: https://europepmc.org/articles/PMC8957912.

[19] Tania Moretta and Giulia Buodo. "Modeling Problematic Facebook Use: Highlighting the role of mood regulation and preference for online social interaction". In: *Addictive Behaviors* 87 (2018), pp. 214–221.

[20] Igor Pantic. "Online social networking and mental health". In: *Cyberpsychology, Behavior, and Social Networking* 17.10 (2014), pp. 652–657.

[21] Mustafa Savci, Ahmet Tekin, and Jon D Elhai. "Prediction of problematic social media use (PSU) using machine learning approaches". In: *Current Psychology* (2020), pp. 1–10.

[22] Maurice H Seevers. "Medical perspectives on habituation and addiction". In: *JAMA* 181.2 (1962), pp. 92–98.

[23] DongBack Seo and Soumya Ray. "Habit and addiction in the use of social networking sites: Their nature, antecedents, and consequences". In: *Computers in Human Behavior* 99 (2019), pp. 109–125.

[24] Yalin Sun and Yan Zhang. "A review of theories and models applied in studies of social media addiction and implications for future research". In: *Addictive Behaviors* 114 (2021), p. 106699. ISSN: 0306-4603. DOI: https://doi.org/10.1016/j.addbeh.2020.106699. URL: https://www.sciencedirect.com/science/article/pii/S0306460320308297.

[25] Monideepa Tarafdar et al. "Explaining the link between technostress and technology addiction for social networking sites: A study of distraction as a coping behavior". In: *Information Systems Journal* 30.1 (2020), pp. 96–124.

[26] M-P Vaillancourt-Morel et al. "For the love of being liked: a moderated mediation model of attachment, likes-seeking behaviors, and problematic Facebook use". In: *Addiction Research & Theory* 28.5 (2020), pp. 397–405.

[27] Chuang Wang and Matthew KO Lee. "Why we cannot resist our smartphones: investigating compulsive use of mobile SNS from a Stimulus-Response-Reinforcement perspective". In: *Journal of the Association for Information Systems* 21.1 (2020), p. 4.

[28] Elisa Wegmann et al. "Interactions of impulsivity, general executive functions, and specific inhibitory control explain symptoms of social-networks-use disorder: An experimental study". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[29] Nikos Xanidis and Catherine M Brignell. "The association between the use of social network sites, sleep quality and cognitive function during the day". In: *Computers in human behavior* 55 (2016), pp. 121–126.