

Incorporating Relapse Rate into Epidemiological Models of Hate Speech Spread

Nitin Agarwal^{1,2}, Hakan Erdem¹

¹COSMOS Research Center, University of Arkansas-Little Rock, USA

²International Computer Science Institute, University of California, Berkeley California, USA
nagarwal@ualr.edu, herdem@ualr.edu

Abstract

Online toxicity spreads through social interactions in ways that resemble contagion processes observed in epidemiology. Prior studies have applied compartmental epidemic models to analyze the propagation of toxic behavior in digital environments, yet these approaches typically assume that recovered individuals must return to a susceptible state before becoming toxic again. In practice, many users rapidly re-engage in toxic activity after temporary recovery. To capture this behavior, we extend the SEIRS epidemiological framework by introducing an immediate relapse mechanism that allows transitions directly from the recovered state to the infected state. Using four datasets spanning political protests, public health debates, and controversial online discussions, we compare the standard SEIRS model with the proposed SEIRS-relapse formulation. Results show that incorporating relapse dynamics reduces L_2 norm error and improves robustness across different sets of parameter initializations. These findings suggest that relapse mechanisms provide a useful extension for modeling recurring toxicity patterns in online discourse.

1 Introduction

In recent years, online toxicity has emerged as a significant challenge affecting social media platforms and online communities. Toxic behavior, characterized by harmful language and hostile interactions, can spread rapidly within digital environments and negatively impact user well-being and the overall quality of online discourse (Sahana et al. 2020). Due to the contagious nature of such behavior, researchers have increasingly drawn parallels between the spread of toxicity and the transmission of infectious diseases (Addai, Yousefi, and Agarwal 2025). This analogy has led to the application of epidemiological models to study and predict the propagation of toxic behavior in social media environments.

Among these models, the SEIRS (Susceptible–Exposed–Infected–Recovered–Susceptible) framework has been used to capture cyclical dynamics of behavioral contagion (Tong et al. 2020). In the context of online toxicity, individuals may transition from being susceptible to exposure through interaction with toxic content, become actively toxic, and eventually recover before becoming susceptible again. However, this formulation assumes that

users must pass through the full susceptible–exposed cycle before becoming toxic again.

Recurring engagement in harmful behavior has also been discussed in research on aggression more broadly. Behavioral neuroscience studies suggest that certain forms of aggression exhibit addiction-like characteristics, including relapse after periods of abstinence (Golden and Shaham 2018). While online toxicity differs from physical aggression, these findings highlight the possibility that harmful behaviors can re-emerge after temporary disengagement. Empirical analyses of toxic interaction networks further indicate that some users repeatedly participate in hostile exchanges and return to toxic discussions over time (Cheng et al. 2017; Kumar et al. 2023). These patterns motivate the inclusion of relapse mechanisms when modeling toxicity propagation dynamics.

To capture this phenomenon, we extend the traditional SEIRS framework by introducing an immediate relapse mechanism that allows individuals to transition directly from the recovered state back to the infected state, which we call SEIRS-relapse. This modification models users who rapidly re-engage in toxic behavior, bypassing the susceptible and exposed stages. Specifically, this study investigates whether incorporating an immediate relapse mechanism into epidemiological models improves their ability to capture toxicity propagation dynamics across multiple online contexts.

We tested and compared the performance of SEIRS and SEIRS-relapse. Our empirical findings from multiple datasets show that SEIRS-relapse improves the robustness of model performance in different parameter initialization scenarios and improves the average performance compared to SEIRS.

The remainder of the paper is organized as follows. Section 2 reviews related work on toxicity and epidemiological modeling in online environments. Section 3 describes the methodology, including data collection, toxicity measurement, and the proposed epidemiological model. Section 4 presents the experimental results. Section 5 discusses the implications of the findings, and Section 6 concludes the study and outlines directions for future research.

2 Related Work

2.1 Toxicity Analysis on Digital Media

The spread of toxicity on social media has become a critical area of research as online platforms increasingly shape public discourse. Researchers have analyzed toxicity in online discussions across different platforms. While some focused on specific events such as the COVID-19 pandemic (DiCicco et al. 2023; Yousefi et al. 2023; Pascual-Ferrá et al. 2021), others have applied computational approaches to detect and analyze toxic behavior on social media platforms including Reddit (Falade, Yousefi, and Agarwal 2024; Almerkhi, Jansen, and Kwak 2020) and X (formerly Twitter) (Vaidya, Nagar, and Nanavati 2024).

A large body of research has focused on detecting and classifying toxic content using machine learning and deep learning approaches (Fan et al. 2021; Taleb et al. 2022; Bonetti et al. 2023; Patel et al. 2024). Beyond detection, several studies have examined behavioral patterns associated with toxic users (Cheng et al. 2017; Chang and Danescu-Niculescu-Mizil 2019). Their findings show that banned users often concentrate their activity within particular discussion threads, post inflammatory or irrelevant content, and provoke responses from other users, contributing to the escalation of toxic interactions.

Together, these studies demonstrate that toxic behavior is not only prevalent in online environments but can also spread through social interactions and user engagement. Although this body of research provides valuable insights into the detection and behavioral dynamics of toxicity, most studies focus on measuring or predicting toxic content rather than modeling the underlying mechanisms that drive its propagation and persistence within online communities.

2.2 Epidemiological Modeling of Information and Behavior

Due to the contagious nature of online interactions, researchers have increasingly drawn parallels between the spread of information and the transmission of infectious diseases. One of the earliest works establishing this analogy is the rumor propagation model proposed by Daley and Kendall (1964), which applied concepts from mathematical epidemiology to describe how information spreads through populations. Subsequent research further extended these ideas to networked systems. For example, Newman (2002) demonstrated that classical epidemiological models such as SIR can be solved exactly on various network structures, highlighting their applicability to complex social systems.

Building on these foundations, epidemiological models have been widely adopted to study diffusion processes in digital environments. Abdullah and Wu (2011) proposed a framework that uses epidemic modeling techniques to analyze the spread of news on Twitter. Similarly, Gardner, Beard, and Medhi (2017) developed the IoT Botnet with Attack Information (IoT-BAI) model, based on a modified SEIRS framework, to study the propagation of malware such as the Mirai worm in IoT networks.

More recent studies have applied epidemiological approaches directly to the study of behavioral dynamics on social media. Nie et al. (2021) incorporated information entropy into a modified SEIR model to analyze the interaction between information dissemination and epidemic transmission. Addai, Yousefi, and Agarwal (2024) introduced a modified SEIQR model that incorporates memory effects to capture how past behavior influences future toxic activity. Additionally, Yousefi and Agarwal (2024) compared several epidemiological models, including SIR, SIS, and STRS, to examine how different model structures capture varying levels of toxicity intensity in online discussions.

While these studies demonstrate the usefulness of epidemiological frameworks for modeling behavioral contagion, existing models focus primarily on transmission and recovery dynamics. The behavior of users who rapidly return to toxic activity after recovery remains relatively underexplored. Addressing this gap is essential for better understanding persistent toxicity cycles in online communities.

3 Methodology

This section details the data collection process, the toxicity detection method, and the explanation and formulation of the epidemiological models.

3.1 Data Collection

To examine toxicity propagation across different types of online discourse, we utilized four datasets representing distinct thematic domains: political protests, public health debates, and controversial online discussions. These datasets were selected to capture diverse contexts in which toxic behavior frequently emerges, and adapted from prior studies or publicly available datasets explained in the following sections. Table 1 summarizes the basic statistics of the datasets used in this study.

Political Protest Domain For the political protest domain, we analyzed the discourse on X surrounding anti-government protests in Brazil following the October 2022 presidential election. Allegations of electoral fraud during this period led to widespread protests, making them a relevant context for studying toxicity propagation in socio-political discussions (Rossini, Mont’Alverne, and Kalogeropoulos 2023). We utilized the dataset from previous studies (Amure and Agarwal 2025; Bhattacharya, Spann, and Agarwal 2024), and refer it as *Brazil Antigov*.

Public Health Debate Domain For the health-related domain, we analyzed the discourse related to opposition to COVID-19 public health measures on X. We utilized the dataset from an earlier study (Yousefi et al. 2023), capturing a prominent phase of global COVID-19 debates on social media. We refer this dataset as *Anti-Covid*.

Controversial Online Discussions The remaining two datasets capture controversial online discussions from Reddit communities where heated debates frequently emerge.

First, we used a publicly available Reddit Climate Change dataset on Kaggle (Lexyr 2022). According to the dataset description from the Kaggle, it was collected to facilitate the

Dataset	Platform	Timeframe	Duration (Days)	Toxic Posts	Size
Brazil Antigov	X	Nov 1 2022 - Jan 30 2023	91	16617	403996
Anti-Covid	X	Mar 12 2020 - Dec 31 2020	295	4420	21031
Climate Change	Reddit	Sep 1 2021 - Sep 12 2022	377	47730	679784
No New Normal	Reddit	Jun 6 2020 - Aug 11 2021	432	3502	17707

Table 1: Statistics of the datasets used in the study. Toxic posts are identified using the toxicity detection procedure described in Section 3.2. *Brazil Antigov* represents the Brazil anti-government dataset from the political protest domain. Similarly *Anti-Covid* represent Covid-19 dataset from public health debate domain. Finally, *Climate Change* and *No New Normal* represents the datasets selected as a part of the controversial online discussions theme.

analysis of climate-related discourse and public debate. We refer this dataset as *Climate Change*.

The second dataset focuses on the *r/NoNewNormal* subreddit, a large Reddit community formed during the COVID-19 pandemic that opposed public health measures. The subreddit grew to over 100,000 members before being quarantined and eventually banned by Reddit in 2021 due to widespread misinformation and coordinated harassment of other communities (Milmo 2021). We utilized the dataset from (Falade, Yousefi, and Agarwal 2024), and it is referred as *No New Normal*.

3.2 Toxicity detection

None of the datasets used in this study contained pre-labeled toxicity annotations. To estimate toxicity levels for each text entry, we employed the Detoxify model (Hanu and Unitary team 2020). Detoxify is a transformer-based model built on the *xlm-roberta-base* architecture and fine-tuned on the Jigsaw Multilingual Toxic Comment Classification dataset¹. It is well established and peer-reviewed in multiple studies (Köpf et al. 2023; Gadre et al. 2023; Falkenberg et al. 2024). The model produces a toxicity score in the range $[0, 1]$, where higher values indicate greater toxic content.

Following a prior study (Saveski, Roy, and Roy 2021), we classify a text as toxic if its toxicity score is greater than or equal to 0.5. This binary classification is used to determine the number of toxic posts reported in Table 1 and to construct the time-series used in the epidemiological models explained in Section 3.7.

3.3 Epidemiological Models

Epidemiological models provide a useful framework for studying how behaviors spread within populations. By dividing a population into compartments such as susceptible, exposed, infected, and recovered, these models allow researchers to analyze how interactions between individuals lead to the propagation of a phenomenon over time.

In this study, we reinterpret epidemiological compartments to reflect user behavior in online discussions. Susceptible (S) users are individuals who have not yet posted toxic content. Exposed (E) users have encountered toxic content and are in a latent stage before potentially engaging in similar behavior. Infected (I) users actively contribute to the spread of toxicity by posting harmful content. Recovered (R)

users represent individuals who have stopped posting toxic content, at least temporarily. This formulation allows us to model how users transition between behavioral states and how toxicity propagates through online communities over time.

3.4 SEIRS Model

The SEIRS model extends classical epidemic models by introducing an exposed compartment that represents a delay between exposure to toxic content and active participation in it. Additionally, the model allows recovered individuals to return to the susceptible state at rate δ , enabling repeated engagement with toxic behavior. The model parameters are defined as follows:

- β – effective contact rate governing transitions from susceptible to exposed
- σ – progression rate from exposed to infected
- γ – recovery rate from infected to recovered
- δ – rate at which recovered users become susceptible again
- λ – recruitment rate representing new users entering the system
- μ – departure rate representing users leaving the discussion

To reflect the dynamic nature of online discussions, we incorporate both a recruitment rate (λ) representing new users entering the system and a departure rate (μ) representing users leaving the discussion. The resulting compartmental structure is illustrated in Figure 1a. These dynamics are represented in the SEIRS model’s ordinary differential equations (ODEs) in Eq. 1, which comprehensively describes the flow of users between states. The equation forms the mathematical foundation for studying the propagation of toxicity across a population, enabling us to quantify how users transition through these stages over time.

$$\text{SEIRS ODEs} \quad \begin{cases} \frac{dS}{dt} = \lambda N(t) + \delta R - \frac{\beta IS}{N(t)} - \mu S, \\ \frac{dE}{dt} = \frac{\beta IS}{N(t)} - (\mu + \sigma)E, \\ \frac{dI}{dt} = \sigma E - (\mu + \gamma)I, \\ \frac{dR}{dt} = \gamma I - (\mu + \delta)R \end{cases} \quad (1)$$

¹<https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge>

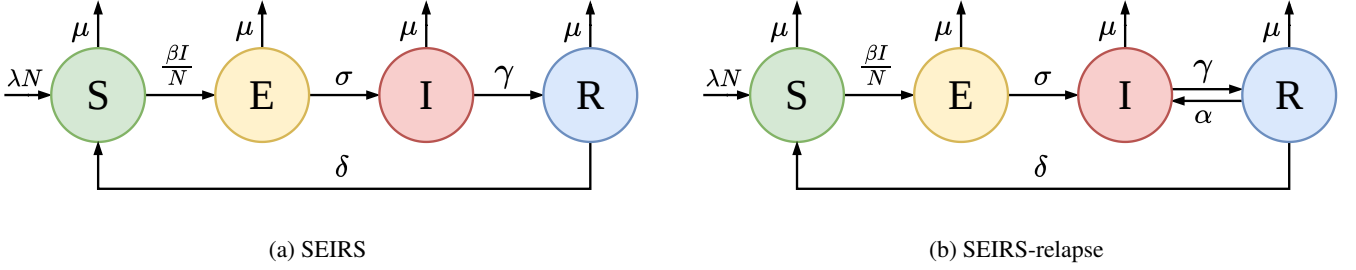


Figure 1: Transfer diagrams for epidemiological models. Figure visualizes the compartments (S)usceptible, (E)xposed, (I)nfectious, and (R)ecovered. Moreover, it shows which parameters govern the transitions between these compartments.

Quantity N from Eq. 1, representing the population count, is defined by Eq. 2 at any given time t .

$$N(t) = S(t) + E(t) + I(t) + R(t). \quad (2)$$

3.5 SEIRS Model with Immediate Relapse

While the SEIRS model captures long-term reinfection through transitions from recovered to susceptible states, it does not explicitly model users who quickly return to toxic behavior shortly after recovering.

To address this limitation, we extend the SEIRS model by introducing an immediate relapse rate α , which represents the direct transition from the Recovered (R) state back to the Infected (I) state. We theorize that the SEIRS model with immediate relapse provides a more accurate depiction of toxicity dynamics by accounting for both short and long term reinfection cycles.

In this modified model, the flow between compartments remains the same, except for the addition of the direct R-to-I pathway governed by the rate α . Figure 1b illustrates the revised SEIRS model. The corresponding dynamics are captured in Eq. 3, which incorporates the new relapse term and updates the overall system of differential equations to reflect this behavior. The changes in the ODEs are highlighted with bold.

$$\text{SEIRS Relapse ODEs} \quad \begin{cases} \frac{dS}{dt} = \lambda N(t) + \delta R - \frac{\beta IS}{N(t)} - \mu S, \\ \frac{dE}{dt} = \frac{\beta IS}{N(t)} - (\mu + \sigma)E, \\ \frac{dI}{dt} = \sigma E - (\mu + \gamma)I + \alpha R, \\ \frac{dR}{dt} = \gamma I - (\mu + \delta)R - \alpha R \end{cases} \quad (3)$$

3.6 Parameter Bounds and Initialization

Parameter Bounds Parameter bounds were chosen to reflect plausible behavioral timescales for toxicity dynamics in online discussions. Since the underlying compartment transitions are not directly observable, these bounds act as guiding constraints that keep the optimization within behaviorally reasonable ranges rather than representing precise empirical estimates. For rate parameters, bounds were

specified by defining a minimum and maximum transition duration and then converting these durations into rates using their reciprocals. For example, a bound of $[1/30, 1/2]$ corresponds to an average transition time between 30 days and 2 days. Table 2 summarizes the bounds and fixed values used for each model parameter.

The bounds are designed to capture relative differences in how quickly different behavioral transitions occur in online discussions. Prior work has shown that toxic behavior in online discussions is often recurrent, with some users repeatedly engaging in such interactions over time and across contexts (Kumar et al. 2023). These observations motivate modeling exposure, infection, and relapse as faster processes compared to recovery and return-to-susceptibility.

The transmission rate β was assigned a relatively wide bound, $[10^{-6}, 10]$, because the intensity of toxicity propagation may vary substantially across datasets and discussion contexts. The exposed-to-infected rate σ was bounded to reflect a short latent period, corresponding to transitions on the order of hours to approximately one day, capturing rapid responses to toxic content. The recovery rate γ was bounded to allow infected users to remain active between 2 and 30 days, reflecting sustained participation in toxic discussions. The return-to-susceptible rate δ was bounded more conservatively, corresponding to a return period between 7 and 180 days, representing slower disengagement and re-entry dynamics.

The recruitment and departure rates, λ and μ , were fixed at 10^{-4} to represent low-rate background turnover in the effective discussion population. These values were kept small because both parameters interact with the total population size and could otherwise dominate the behavioral dynamics. Fixing λ and μ also reduces model flexibility and helps stabilize parameter estimation across datasets and model variants.

Finally, the relapse rate α was allowed to vary within the range $[10^{-6}, 1/3]$, corresponding to relapse timescales ranging from very infrequent recurrence to rapid re-engagement on the order of a few days.

Initialization In addition to defining parameter bounds, careful consideration was given to the initialization of model parameters during optimization. Nonlinear optimization of epidemiological models can be sensitive to initial parameter values and may converge to different local minima depend-

Parameter	Bound/Value	Description
β	$[10^{-6}, 10]$	transmission rate
γ	$[1/30, 1/2]$	recovery rate
σ	$[1, 3]$	exposed \rightarrow infected
δ	$[1/180, 1/7]$	recovered \rightarrow susceptible
λ	10^{-4}	recruitment rate (fixed)
μ	10^{-4}	departure rate (fixed)
α	$[10^{-6}, 1/3]$	relapse rate

Table 2: Bounds of the parameters that govern the transitions between compartments. We selected the bounds for each parameter heuristically. A bound of $[1/180, 1/7]$ means the range for that parameter is 7 to 180 days. That is the the reciprocal of the selected bounds.

ing on the starting point. To mitigate this issue, we adopted a multi-start initialization strategy in which multiple parameter configurations are used as starting points for model fitting.

Initial parameter vectors were generated using Latin hypercube sampling (LHS), a sampling technique designed to efficiently cover multidimensional parameter spaces. LHS ensures that sampled parameter values span the entire range of each bound while avoiding clustering of samples in specific regions of the search space (Helton and Davis 2003). Using this approach, we generated 100 distinct parameter initializations covering the feasible parameter domain.

To ensure a fair comparison between the SEIRS and SEIRS-relapse models, both models shared identical initial values for their common parameters ($\beta, \gamma, \sigma, \delta, \lambda, \mu$). For the SEIRS-relapse model, the additional relapse parameter α was sampled independently within its specified bound and appended to the corresponding base parameter vector. The resulting parameter sets were saved and reused across all datasets to ensure consistent experimental conditions. Each parameter configuration served as the starting point for a separate optimization run, producing a total of 100 model fits per dataset for each model variant.

3.7 Construction of the Toxicity Time Series

After assigning toxicity scores to each post, we constructed a temporal series representing the evolution of toxic activity in each dataset. In our experimental setup, we approximate the infected population using the number of toxic posts due to the absence of user-level behavioral states. This design choice enables this framework to be applied across datasets, but we still acknowledge its simplistic nature. The infected time series $I(t)$ was obtained by aggregating the number of toxic posts within fixed time intervals.

Specifically, for each dataset we counted the number of toxic posts per day:

$$I(t) = \text{number of toxic posts observed at day } t$$

This series serves as the empirical observation of the infected compartment in the epidemiological model. While the SEIRS framework includes additional compartments ($S, E,$ and R), these states are not directly observable from the data

and are inferred indirectly through the model dynamics during parameter estimation.

3.8 Model Fitting

To evaluate the ability of the epidemiological models to capture toxicity dynamics, model parameters were estimated by fitting the simulated infected trajectory to the observed toxicity time series.

Effective Population Size The total population size N required by the compartmental model is not directly observable from the toxicity datasets. Therefore, in our setting, N should not be interpreted as the true number of unique participants, but rather as an *effective discussion population* that sets the scale of the compartmental dynamics, particularly the susceptible pool and the transmission term.

To obtain a stable and representative estimate of this scale, we derive N from the infected time series prior to model fitting. Specifically, we use a high-percentile statistic of the observed infected counts:

$$N = 3 \times Q_{95}(I) \quad (4)$$

The 95th percentile is used as a stable approximation of peak activity. Instead of relying on the maximum value in the observed data, which may reflect the characteristic of a single day, we use a high percentile to capture typical high-activity periods.

The multiplicative factor is a heuristic scaling constant used to expand the observed toxic activity to a larger discussion population. Since not all participants produce toxic content, the number of toxic posts reflects only a portion of overall activity. The factor of three is not intended to represent a precise proportion of toxic to non-toxic users, but rather to ensure that the effective population size is sufficiently larger than the observed infected counts to support meaningful compartmental dynamics. In this sense, it serves as a practical scaling choice rather than an empirically derived parameter.

We note that this estimation is heuristic and primarily serves to initialize the model and scale the interaction dynamics. While different choices of N may lead to variations in the estimated parameter values, our objective is to compare model structures under a consistent setup. Since both SEIRS and SEIRS-relapse models are fit using the same N for each dataset, the relative comparison between models remains unaffected by this choice.

Given the estimated population size N , the model compartments are initialized at the beginning of each simulation. The infected compartment is initialized using the first observed toxicity value. Because the exposed and recovered states are not directly observable in the data, they are initialized to zero. The remaining population is assigned to the susceptible compartment:

$$\begin{aligned} I_0 &= I(t_0) \\ S_0 &= N - I_0. \\ E_0 &= 0 \\ R_0 &= 0 \end{aligned} \quad (5)$$

This initialization assumes that toxicity propagation begins with the observed infected users while the remainder of the discussion population is initially susceptible.

Parameter Estimation Parameter estimation was performed using bounded optimization (L-BFGS-B) with numerical integration of the ODE system using a standard ODE solver. For each of the 100 initializations described in the previous subsection, bounded optimization using the bounds in Table 2 is performed to identify parameter values that best reproduce the observed dynamics. Model performance is evaluated using the relative L_2 error (Eq. 6) between the simulated infected trajectory and the observed toxicity time series.

$$L_2 \text{ Error} = \frac{\|I_{\text{model}}(t) - I(t)\|_2}{\|I(t)\|_2}. \quad (6)$$

This metric measures the overall discrepancy between the model prediction and the observed toxicity dynamics across the full time series. A lower error value indicates a closer match between the simulated and observed trajectories.

4 Results and Findings

We applied both the SEIRS and SEIRS-relapse models for 100 trials on each dataset using different parameter initializations. Table 3 presents the average error rates obtained for each dataset and model. Across all datasets, the SEIRS-relapse model achieves lower error rates than the standard SEIRS model. The magnitude of improvement varies between datasets. The greatest error reduction is observed in the Brazil Antigov dataset, where the relapse model improves the error rate by 4.69%. In contrast, the improvement is more modest in the No New Normal dataset, where the reduction is 0.39%. These results indicate that incorporating relapse dynamics improves the model’s ability to reproduce observed toxicity patterns.

Figure 2 provides a more detailed view of the distribution of error rates across the 100 trials for each dataset. The box plots and cumulative distribution functions (CDF) reveal consistent differences between the models. In particular, the relapse model exhibits lower standard deviations of error across most datasets, suggesting that the additional relapse parameter improves the stability of the model under different parameter initializations.

The Climate Change dataset presents a slightly different pattern. While the standard SEIRS model occasionally achieves lower minimum error values in lower percentiles, the SEIRS-relapse model performs better across the majority of parameter initializations. In other words, SEIRS is capable of reaching slightly better best-case fits, but SEIRS-relapse produces consistently lower errors across the full distribution of trials. This suggests that the relapse model provides greater robustness to parameter initialization and reduces the likelihood of poor local optima during optimization.

To further examine model behavior, we analyzed the predicted infected trajectories and compared them with the observed toxicity time series. Figure 3 shows the average predicted $I(t)$ trajectories across all trials alongside the ob-

Dataset	SEIRS	SEIRS-relapse	Diff.
Brazil Antigov	67.76	63.07	4.69
Anti-Covid	50.17	48.93	1.24
Climate Change	41.67	40.39	1.28
No New Normal	56.07	55.68	0.39

Table 3: Average L_2 norm error rates (%) for SEIRS and SEIRS-relapse models across parameter initialization trials. Brazil Antigov experiments showed the biggest improvement with 4.69% change in L_2 norm error on average.

served daily toxic post counts. The differences between the two models are most pronounced in the Brazil Antigov dataset, where the SEIRS-relapse model more closely follows the peaks and troughs of the observed data. For the remaining datasets, the trajectories of the two models are more similar, which is consistent with the smaller improvements observed in the error metrics.

Finally, we analyzed the distribution of the relapse parameter α across all trials for each dataset (Figure 4). The model fits frequently depict α above the lower bound of 10^{-6} . In the Brazil Antigov dataset, approximately 90% of trials estimate α above the lower bound, indicating that the relapse pathway is actively utilized in the majority of fitted solutions. For the remaining datasets, roughly 75% to 80% of trials estimate α above the lower bound, suggesting that relapse dynamics are also frequently incorporated into the fitted models, though somewhat less consistently than in the Brazil dataset. This difference is also evident in Figure 3, where the Brazil dataset shows more dynamic fits produced by the relapse model.

In addition to approximately 20% to 25% of trials across several datasets converging to the lower bound, a substantial fraction of estimates in all datasets except Climate Change accumulate near the upper bound of $\alpha = \frac{1}{3}$, accounting for roughly 35% to 40% of the parameter estimates. This concentration of solutions near both bounds suggests that the current parameter range may restrict the optimization process in some cases. Future work may therefore explore wider bounds for α to allow the model to capture a broader range of relapse dynamics.

5 Discussion

From a methodological perspective, these findings indicate that extending epidemiological models with relapse dynamics potentially offer a useful approach for studying recurring behavioral phenomena in online discourse. Toxicity in social media environments often emerges in bursts, declines, and then reappears over time. The relapse mechanism allows the model to capture these recurring patterns more effectively than classical SEIRS dynamics alone. While our results remain empirical and do not imply that relapse extensions will always outperform baseline models, they suggest that incorporating relapse dynamics can improve model stability and representational capacity in many cases.

More broadly, this study highlights the value of relapse as a conceptual component in computational models of online toxicity. By formalizing relapse within a compartmental

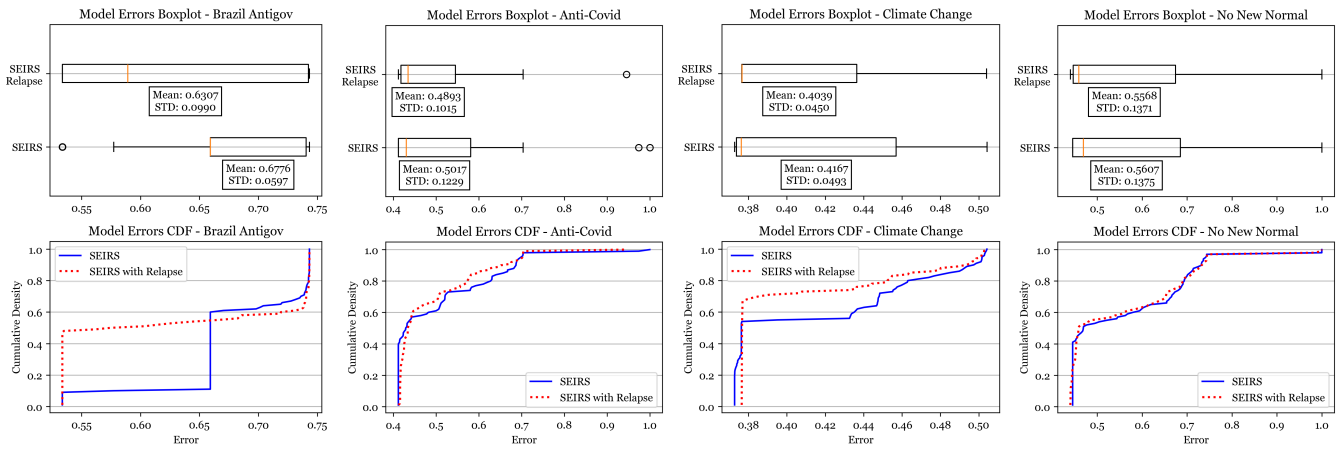


Figure 2: Distribution of L_2 norm errors across all parameter initialization trials. Figure shows the boxplot distribution of the trials with the mean and standard deviation values reported. Moreover, it shows the cumulative distribution function (CDF) of the trials, providing a different lens for the comparison of trial results. In simple terms for CDF, the more error rates closer to the upper left the more robust the model with different parameter initializations.

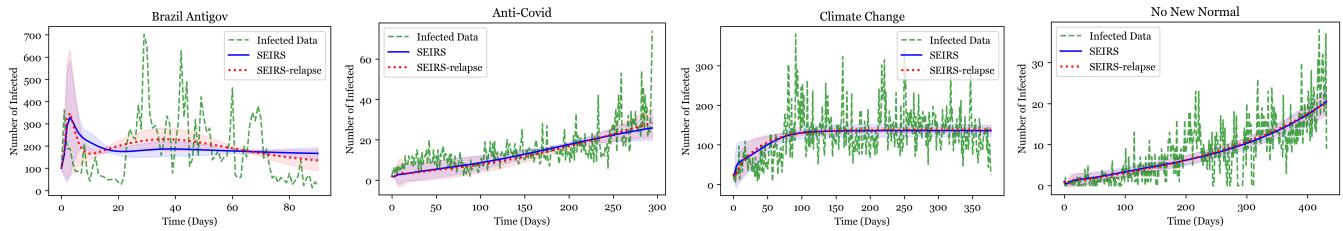


Figure 3: Average simulated trajectories of the I compartment by the models compared with observed daily toxic post counts. The red and blue lines show the model estimations, and green dashed line shows the observed data. The shades behind the lines show the standard deviation of the averaged estimations.

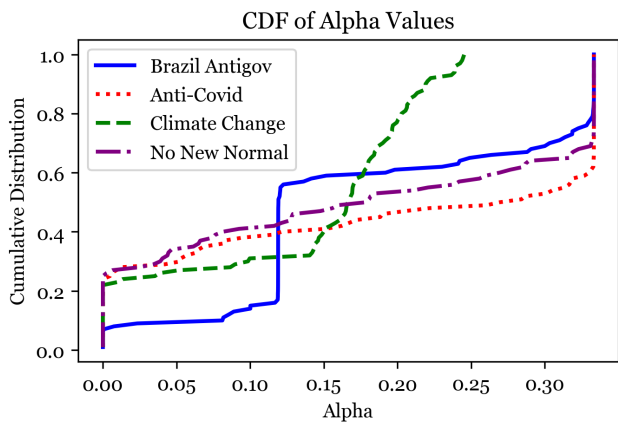


Figure 4: CDF visualization of estimated α values of the SEIRS-relapse model for each dataset. In each dataset, the usage profile of α shows differences. Moreover the clustered mass in the lower and upper bounds show that wider limits for α should be considered in future studies.

modeling framework, we demonstrate how recurring toxic activity can be incorporated into quantitative analyses of discourse dynamics. Although our model does not directly observe user-level behavior, the relapse mechanism provides a useful abstraction for representing repeated toxicity activity over time. This concept may also be incorporated into other computational approaches beyond epidemiological modeling, such as machine learning systems that analyze temporal patterns of toxic discourse.

Our findings also suggest that toxicity dynamics vary across different types of online discussions. The distribution of the relapse parameter α differed across datasets, indicating that the relative importance of relapse dynamics varies depending on the context of the discourse. This observation suggests that flexible modeling frameworks may be necessary to capture the diverse propagation patterns observed in different online communities.

Finally, epidemiological models provide a framework for analyzing propagation dynamics in online discourse. Once the parameters are estimated, these models can be used to derive quantities such as the basic reproduction number (R_0), which describes the expected growth potential of a phenomenon within a population. Although the present study focuses on improving model fitting rather than ana-

lyzing these derived quantities in depth, future research may use relapse-based epidemiological models to calculate the intensity of propagation of toxic discourse over time.

6 Conclusion, Limitations, and Future Work

This study investigated whether incorporating relapse dynamics into epidemiological models improves the ability to capture toxicity propagation in online discussions. Across multiple datasets, the SEIRS-relapse model consistently achieved lower average error rates compared to the standard SEIRS formulation and demonstrated improved robustness across distinct parameter initializations. These results suggest that relapse dynamics provide a useful extension for modeling recurring toxic activity in online discourse. While the improvements observed in this study remain empirical and dataset-dependent, they indicate that incorporating relapse mechanisms can increase the flexibility of epidemiological models when fitting toxicity trajectories.

Beyond the empirical findings, several methodological considerations and limitations should be acknowledged when interpreting the results. First, the experimental setup does not model individual users directly. Instead, toxic posts are used as a proxy for infected individuals, meaning that the infected compartment represents aggregated toxic activity rather than user-level behavioral states. As a result, the model captures patterns of toxicity production rather than transitions of individual users between behavioral states.

Second, the available data only provide observations for the infected compartment $I(t)$, while the susceptible, exposed, and recovered compartments remain unobserved. This partial observability introduces parameter identifiability challenges, as multiple parameter combinations may produce similar infected trajectories. Consequently, estimated parameters should be interpreted as effective parameters describing the observed dynamics rather than precise behavioral transition rates. This limitation also affects the stability of derived quantities such as the basic reproduction number (R_0), which may exhibit sensitivity to parameter uncertainty. Therefore, any further study that investigates the rate of propagation should consider providing additional compartment observations to the models.

Third, the parameter bounds used during optimization were selected heuristically rather than derived from established theoretical or empirical estimates of toxicity dynamics. The bounds were intended to restrict the search space to behaviorally plausible timescales for exposure, recovery, and relapse. However, because these ranges are not grounded in validated behavioral measurements, the fitted parameters potentially reflect the imposed search space rather than the actual underlying behavioral mechanisms.

Fourth, the study employs a population-level compartmental model that assumes homogeneous mixing among participants in the discussion. In reality, online discourse occurs within complex network structures where interactions are shaped by follower networks, reply chains, and platform algorithms. The present framework abstracts away these structural effects in order to operate on datasets where reliable interaction networks is not always be available.

Finally, while the relapse model demonstrated improvements across the datasets presented in this study, the results should not be interpreted as evidence that relapse dynamics will universally outperform classical epidemiological models. Toxicity propagation patterns vary across online communities and discussion contexts, and the relative usefulness of relapse mechanisms depend on the temporal characteristics of the discourse being modeled.

Future work may address these limitations by incorporating user-level behavioral trajectories, integrating network structures into the modeling framework, and exploring alternative formulations of relapse dynamics. In addition, expanding the analysis to larger collections of datasets and exploring wider and empirically grounded parameter ranges may further clarify the conditions under which relapse dynamics provide meaningful improvements for modeling toxicity propagation.

Statements and Declarations

Competing interest: The writers state that they do not have any competing interests.

Data availability statement: The data supporting the conclusions of this study will be made available upon request following the terms and conditions of the social media platforms.

Use of generative AI in scientific writing: Generative AI and AI-assisted technologies were not used in the preparation of this manuscript.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

- Abdullah, S.; and Wu, X. 2011. An Epidemic Model for News Spreading on Twitter. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, 163–169.
- Addai, E.; Yousefi, N.; and Agarwal, N. 2024. SEIQR: An Epidemiological Model to Contain the Spread of Toxicity

- using Memory-Index. In *Proceedings of the Fifth International Workshop on Cyber Social Threats (CySoc 2024)*, ICWSM 2024 Workshops. Buffalo, New York, USA.
- Addai, E.; Yousefi, N.; and Agarwal, N. 2025. Utilizing Fractional Order Epidemiological Model to Understand High and Moderate Toxicity Spread on Social Media Platforms. In *Social Networks Analysis and Mining*, 298–308. Cham: Springer Nature Switzerland. ISBN 978-3-031-78538-2.
- Almerexhi, H.; Jansen, B. J.; and Kwak, H. 2020. Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators. In *Companion proceedings of the Web Conference 2020*, WWW '20, 294–298. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370240.
- Amure, R.; and Agarwal, N. 2025. Modeling Word-Level Functions in Social Movement Discourse. In *ICIS 2025 Proceedings*, number 30 in User Behavior.
- Bhattacharya, S.; Spann, B.; and Agarwal, N. 2024. A Computational Approach to Analyze Identity Formation: A Case Study of Brazil Insurrection. In *Proceedings of the Americas Conference on Information Systems (AMCIS 2024)*, 19.
- Bonetti, A.; Martínez-Sober, M.; Torres, J. C.; Vega, J. M.; Pellerin, S.; and Vila-Francés, J. 2023. Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. *Applied Sciences*, 13(10).
- Chang, J.; and Danescu-Niculescu-Mizil, C. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *The World Wide Web Conference*, WWW '19, 184–195. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.
- Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1217–1230. New York, NY, USA: Association for Computing Machinery. ISBN 9781450343350.
- Daley, D.; and Kendall, D. 1964. Epidemics and Rumours. *Nature*, 204: 1118.
- DiCicco, K.; Bin Noor, N.; Yousefi, N.; Spann, B.; and Agarwal, N. 2023. Toxicity and Networks of COVID-19 Discourse Communities: A Tale of Two Media Platforms. In *Proceedings of the 3rd International Workshop on Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2023) co-located with the 45th European Conference on Information Retrieval (ECIR 2023)*. Dublin, Ireland.
- Falade, T. C.; Yousefi, N.; and Agarwal, N. 2024. Toxicity Prediction in Reddit. In *Proceedings of the Americas Conference on Information Systems (AMCIS 2024)*, 18.
- Falkenberg, M.; Zollo, F.; Quattrociocchi, W.; Pfeffer, J.; and Baronchelli, A. 2024. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nature Communications*, 15(1): 9560.
- Fan, H.; Du, W.; Dahou, A.; Ewees, A. A.; Yousri, D.; Elaziz, M. A.; Elsheikh, A. H.; Abualigah, L.; and Alqaness, M. A. A. 2021. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics*, 10(11).
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; Orgad, E.; Entezari, R.; Daras, G.; Pratt, S.; Ramanujan, V.; Bitton, Y.; Marathe, K.; Mussmann, S.; Vencu, R.; Cherti, M.; Krishna, R.; Koh, P. W. W.; Saukh, O.; Ratner, A. J.; Song, S.; Hajishirzi, H.; Farhadi, A.; Beaumont, R.; Oh, S.; Dimakis, A.; Jitsev, J.; Carmon, Y.; Shankar, V.; and Schmidt, L. 2023. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 27092–27112. Curran Associates, Inc.
- Gardner, M. T.; Beard, C.; and Medhi, D. 2017. Using SEIRS Epidemic Models for IoT Botnets Attacks. In *DRCN 2017 - Design of Reliable Communication Networks; 13th International Conference*, 1–8.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Golden, S. A.; and Shaham, Y. 2018. Aggression Addiction and Relapse: A New Frontier in Psychiatry. *Neuropsychopharmacology*, 43(1): 224–225.
- Hanu, L.; and Unitary team. 2020. Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Helton, J.; and Davis, F. 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81(1): 23–69.
- Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; ES, S.; Suri, S.; Glushkov, D.; Dantuluri, A.; Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In *Advances in Neural Information Processing Systems*, volume 36, 47669–47681. Curran Associates, Inc.
- Kumar, D.; Hancock, J.; Thomas, K.; and Durumeric, Z. 2023. Understanding the Behaviors of Toxic Accounts on Reddit. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2797–2807. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.
- Lexyr. 2022. The Reddit Climate Change Dataset. Kaggle. <https://www.kaggle.com/datasets/pavellexyr/the-reddit-climate-change-dataset>.
- Milmo, D. 2021. Reddit Communities Go Dark in Protest over Covid Misinformation. *The Guardian*. <https://www.theguardian.com/technology/2021/sep/01/reddit-communities-go-dark-in-protest-over-covid-misinformation>.
- Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Phys. Rev. E*, 66: 016128.

Nie, Q.; Liu, Y.; Zhang, D.; and Jiang, H. 2021. Dynamical SEIR Model With Information Entropy Using COVID-19 as a Case Study. *IEEE Transactions on Computational Social Systems*, 8(4): 946–954.

Pascual-Ferrá, P.; Alperstein, N.; Barnett, D. J.; and Rimal, R. N. 2021. Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. *Big Data & Society*, 8(1): 20539517211023533.

Patel, D.; Pramanik, P. K. D.; Suryawanshi, C.; and Pareek, P. 2024. Detecting toxic comments on social media: an extensive evaluation of machine learning techniques. *Journal of Computational Social Science*, 8(1): 20.

Rossini, P.; Mont’Alverne, C.; and Kalogeropoulos, A. 2023. Explaining Beliefs in Electoral Misinformation in the 2022 Brazilian Election: The Role of Ideology, Political Trust, Social Media, and Messaging Apps. *Harvard Kennedy School (HKS) Misinformation Review*, 4(3).

Sahana, B.; Sandhya, G.; Tanuja, R.; Ellur, S.; and Ajina, A. 2020. Towards a Safer Conversation Space: Detection of Toxic Content in Social Media (Student Consortium). In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 297–301.

Saveski, M.; Roy, B.; and Roy, D. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021, WWW ’21*, 1086–1097. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Taleb, M.; Hamza, A.; Zouitni, M.; Burmani, N.; Lafkiar, S.; and En-Nahnahi, N. 2022. Detection of toxicity in social media based on Natural Language Processing methods. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1–7.

Tong, Q.; Wang, H.; Zhang, J.; Li, L.; and Huang, Q. 2020. The Fractional SEIRS Epidemic Model for Information Dissemination in Social Networks. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, 284–291. Cham: Springer International Publishing. ISBN 978-3-030-32591-6.

Vaidya, A.; Nagar, S.; and Nanavati, A. A. 2024. Analysing the Spread of Toxicity on Twitter. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD), CODS-COMAD ’24*, 118–126. New York, NY, USA: Association for Computing Machinery. ISBN 9798400716348.

Yousefi, N.; and Agarwal, N. 2024. Study the Influence of Toxicity Intensity on Its Propagation Using Epidemiological Models. In *Proceedings of the Americas Conference on Information Systems (AMCIS 2024)*, 17.

Yousefi, N.; Bin Noor, N.; Spann, B.; and Agarwal, N. 2023. Towards Developing a Measure to Assess Contagiousness of Toxic Tweets. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media (ICWSM 2023): TrueHealth 2023 Workshop on Combating Health Misinformation for Social Wellbeing*.

6.1 Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, proposed model aims to understand the toxicity propagation in digital media rather than to moderate or censor online toxicity. The datasets used do not involve individually identifiable information.**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **N/A. The datasets used are collections of public social media posts filtered by topic-relevant hashtags and keywords. The study does not make claims about population-representative distributions and instead treats each dataset as a self-contained discussion community.**
- (e) Did you describe the limitations of your work? **Yes, in detail**
- (f) Did you discuss any potential negative societal impacts of your work? **No. The work is focused on improving epidemiological model fit for toxicity data. The authors do not foresee direct negative societal impacts. The model does not produce outputs that could be used to target, surveil, or harm individuals.**
- (g) Did you discuss any potential misuse of your work? **No. The epidemiological framework proposed here estimates population-level propagation parameters from aggregated data and is not designed for individual-level targeting.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes. The modeling framework operates exclusively at the population level and does not store or expose individual user identities.**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
- (b) Have you provided justifications for all theoretical results? **NA**
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**

- (e) Did you address potential biases or limitations in your theoretical framework? *NA*
 - (f) Have you related your theoretical results to the existing literature in social science? *NA*
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? *NA*
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? *NA*
 - (b) Did you include complete proofs of all theoretical results? *NA*
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? *NA*
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? *NA*
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? *NA*
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? *NA*
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? *NA*
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? *NA*
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- (a) If your work uses existing assets, did you cite the creators? *Yes*
 - (b) Did you mention the license of the assets? *NA*
 - (c) Did you include any new assets in the supplemental material or as a URL? *NA*
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? *NA*
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? *NA*
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? *NA*
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? *NA*
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? *NA*
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? *NA*
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? *NA*
 - (d) Did you discuss how data is stored, shared, and de-identified? *NA*