Modeling Quarantine Intervention for Varied Toxic Intensities

Nitin Agarwal
COSMOS Research Center, University of Arkansas, Little Rock, USA
International Computer Science Institute, University of California, Berkeley, USA
e-mail: nxagarwal@ualr.edu

***Abstract*—This study employs a Susceptible-Exposed-Infected-Quarantine-Recovered (SEIQR) epidemiological framework to analyze the spread of toxicity in online environments, integrating toxicity intensity stratification to capture the complexity of toxicity propagation. Using datasets from coronavirus disease of 2019 (COVID-19) and social movement discussions, we conduct sensitivity analysis to evaluate key parameters influencing toxicity diffusion. The results reveal that splitting toxicity into moderate and high levels significantly reduces model error rates, enhancing predictive accuracy across all datasets. Additionally, our findings indicate that the basic reproduction number (R_0) is highly sensitive to exposure and quarantine rates, emphasizing the critical role of enhanced moderation and adaptive quarantining in suppressing toxicity. Moreover, quarantine interventions and content demotion strategies are shown to significantly curb toxicity intensity while maintaining engagement dynamics. These insights provide a foundation for policy-driven interventions, enabling social media platforms to implement optimized content moderation, algorithmic intervention, and network-level strategies to mitigate online toxicity and promote healthier digital ecosystems.**

Keywords-toxicity; social media; epidemiological modeling; hate speech; COVID-19; Brazil; Peru.

I. Introduction

The rapid evolution of social media platforms has transformed the way users engage in communication, share information, and participate in public discourse. These platforms serve as catalysts for the swift dissemination of news, educational resources, and discussions on critical societal matters [1]. By enabling real-time interaction across diverse audiences, social media fosters global connectivity and awareness. However, alongside these beneficial aspects, social media has also unintentionally become a breeding ground for the amplification of toxic behaviors. This toxicity manifests in various forms, including hate speech, misinformation, harassment, cyberbullying, and online extremism. The unrestricted and algorithm-driven nature of social platforms often facilitates the viral spread of harmful content, influencing not only individual users but also wider societal dynamics [2]–[5].

The proliferation of toxicity on digital platforms has far-reaching consequences that extend beyond online environments, affecting psychological well-being, social cohesion, and even real-world actions. Misinformation-driven narratives have contributed to public unrest, political polarization, and radicalization, demonstrating the tangible effects of unchecked digital toxicity [6]. In extreme cases, the rapid circulation of false information has led to public health crises, economic disruptions, and coordinated disinformation campaigns aimed at manipulating public opinion [7][8]. As a result, addressing the spread and impact of online toxicity has become an urgent research priority that demands multidisciplinary approaches and advanced analytical frameworks [9][10].

Given the increasing reliance on digital communication and algorithmic content curation, it is imperative to understand the underlying mechanisms that drive toxicity propagation across interconnected online communities. The primary challenge lies in developing intervention strategies that effectively mitigate toxic interactions while upholding fundamental rights to freedom of speech and avoiding excessive restrictions on public discourse. Striking a balance between content regulation and open dialogue is crucial to maintaining healthy digital ecosystems that encourage constructive engagement.

The complexity and scale of modern social networks necessitate the adoption of computational modeling and data-driven methodologies to analyze the spread, persistence, and impact of online toxicity. By leveraging mathematical modeling, network analysis, and artificial intelligence, researchers can identify key transmission patterns, predict emerging toxicity trends, and develop targeted interventions that curb digital toxicity while preserving online freedoms. These insights will help policymakers, social media platforms, and researchers formulate evidence-based strategies to foster safer and more inclusive online environments.

This study addresses the following key research questions:

i **RQ1:** How do variations in toxicity intensity (moderate vs. high vs. no split) influence the impact on the Susceptible-Exposed-Infected-Quarantine-Recovered (SEIQR) model's performance?

ii **RQ2:** What are the key parameters in the SEIQR model that have the most significant impact on the *basic reproduction number* R_0, and how can these parameters be controlled to mitigate toxicity spread?

The remainder of the paper is organized as follows: Section 2 provides a review of related work on toxicity propagation in social media and epidemiological modeling techniques. Section 3 introduces the methodology, detailing the SEIQR model formulation. Section 4 presents the experimental results and discusses the impact of different toxicity intensities. Section 5 presents the results and discussion. Finally, Section 6 concludes the study with recommendations for future research and policy interventions.

II. Related Work

The rise of online toxicity in social media environments has prompted extensive research into its dynamics, effects, and mitigation strategies. With the increasing spread of harmful

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

or toxic content, researchers have employed various computational, epidemiological, and network-based models to understand and control the propagation of toxicity.

A. Machine Learning Approaches to Toxicity Detection

Several studies have leveraged machine learning techniques to detect, classify, and analyze toxicity in online discussions. For example, researchers in [11][12] analyzed coronavirus disease of 2019 (COVID-19) misinformation and toxicity across social media platforms, revealing notable differences in the extent and nature of toxic behaviors on Twitter, Reddit, and Facebook. Their findings emphasized the role of key user groups (super-spreaders) in amplifying misinformation and toxic content. Similarly, natural language processing models such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformers (GPT), and Detoxify have been used to detect toxicity in real time [13]–[16]. These models have been instrumental in the identification of various forms of toxic language, including hate speech, harassment, cyberbullying, and misinformation. Furthermore, studies in [17] investigated the dynamics of discussion threads on Reddit, where toxicity spreads hierarchically within nested conversations. However, these machine learning-based approaches largely focus on static classification of toxicity, often overlooking long-term propagation effects and feedback loops that sustain toxic environments.

B. Epidemiological Models for Online Toxicity Spread

Epidemiological modeling has proven to be an effective framework for understanding and predicting the spread of online toxicity by drawing analogies between toxic content dissemination and infectious disease transmission. Traditional models such as Susceptible-Infected-Recovered (SIR), Susceptible-Infected-Susceptible (SIS), and Susceptible-Exposed-Infected-Recovered (SEIR) have been adapted to analyze how toxicity spreads within digital communities [18][19]. The Susceptible-Toxic-Recovered-Susceptible (STRS) model has further refined this analogy by incorporating recovery mechanisms where users disengage from toxic interactions [10][20][21].

Moreover, quarantine-based control strategies have been widely adopted in epidemiological models to curb infectious spread, an approach that has direct applications in content moderation and toxicity mitigation. Authors in [22] studied the impact of content quarantining on social media toxicity, demonstrating that isolating toxic users reduces the overall infection rate in a network-based toxicity model. Similarly, [20] applied epidemiological modeling to study COVID-19 misinformation spread, highlighting the role of user quarantine, moderation interventions, and algorithmic content suppression.

III. Methodology

This section outlines the data collection framework, analytical methods, and computational procedures employed to validate the proposed fractal-fractional SEIQR model.

A. Data Collection and Analysis

To validate the model, we analyzed two distinct datasets: (1) discourse related to the COVID-19 pandemic (spanning February 2020–June 2021) and (2) social movements in Brazil (spanning November 1, 2022–February 25, 2023) and Peru (spanning December 7, 2022–January 31, 2023). Public posts were collected via X's (formerly Twitter) Academic API, focusing on keywords and hashtags associated with polarized discourse.

- **COVID-19 Dataset**: Included topics such as pandemic policies, face mask mandates, lockdowns, and 5G conspiracy theories. Key hashtags: #fckthecovid/s, #fckyourmask/s, #f*cklockdown/s, #5GCoronavirus.
- **Brazilian Protests**: Centered on post-election unrest following the October 30, 2022, presidential election, with hashtags like #semanistia and #SOSbrasil reflecting demands for military intervention and counter-protests.
- **Peruvian Protests**: Captured anti-government demonstrations after President Pedro Castillo's removal on December 7, 2022, using hashtags such as #peruprotest/s.

1) Toxicity Classification: Toxicity scores for each post were computed using Detoxify [23], a pre-trained deep learning model that evaluates text for harmful content. Detoxify employs convolutional neural networks and semantic embeddings to assign a toxicity probability between 0 (non-toxic) and 1 (highly toxic). Posts with scores of 0.5 or more were classified as toxic; those below 0.5 were deemed non-toxic.

To analyze toxicity intensity, the toxic subset was further stratified:

- **Moderate Toxicity**: Posts with scores below the dataset-specific average toxicity score.
- **High Toxicity**: Posts exceeding the average score.

Table I summarizes the distribution of high/moderate toxic posts and average scores across datasets.

2) User Activity and Quarantining: To identify superpropagators, we isolated the top 10% of users by activity level, defined as the number of retweets per post [22]. These high-engagement users were algorithmically transferred to a quarantine compartment in the SEIQR model, simulating platform-level interventions to curb toxicity spread.

B. Model Formulation

Online toxicity is a growing epidemic on digital platforms, characterized by high transmission rates, latent behavior, and infectiousness. To analyze this, we used the SEIQR model (illustrated in Figure 1) as the following system of ordinary differential equations.

$$\begin{cases} \frac{dS(t)}{dt} = \mathcal{A} + \eta R - \frac{\beta I S}{N(t)} - \mu S, \\ \frac{dE(t)}{dt} = \frac{\beta I S}{N(t)} - (\mu + \psi)E, \\ \frac{dI(t)}{dt} = \psi E - (\mu + \phi + \theta)I, \\ \frac{dQ(t)}{dt} = \theta I - (\gamma + \mu)Q, \\ \frac{dR(t)}{dt} = \gamma Q + \phi I - (\mu + \eta)R, \end{cases} \tag{1}$$

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

TABLE I. Toxicity dataset statistics.

Dataset	Num. of Posts	Avg. Toxicity	Num. of High Toxic	Num. of Moderate Toxic
F*Covid	28,131	0.91	4,684	2,082
F*Mask	2,423	0.91	538	217
F*Lockdown	1,995	0.82	598	493
5G	33,403	0.84	1,096	703
Brazil Anti	405,160	0.70	1,221	2,309
Brazil Pro	44,415	0.75	105	131
Peru	195,290	0.71	511	546

where $\lambda = \frac{\beta IS}{N(t)}$, the entire population, and we define the quantity $N(t)$ by

$$N(t) = S(t) + E(t) + I(t) + Q(t) + R(t).$$

and initial conditions

$$S(0) = S_0 \geq 0, \quad E(0) = E_0 \geq 0, \quad I(0) = I_0 \geq 0,$$
$$Q(0) = Q_0 \geq 0, \quad R(0) = R_0 \geq 0.$$

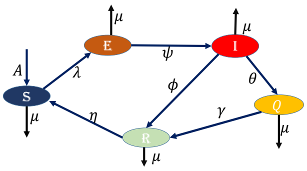

Figure 1. Transfer diagram for the toxicity spread on the social network platform.

IV. Model Parameterization and Data Fitting Analysis

Ensuring accurate model validation and precise parameter estimation is fundamental in mathematical modeling when working with real-world data (illustrated in Table II). The main challenge lies in determining the most suitable parameter values from empirical data, making parameter fitting a crucial step in model formulation. One widely used approach for parameter estimation in nonlinear models is the Non-Linear Least Squares Method (NLSM), which minimizes the discrepancy between observed and predicted values.

Consider a nonlinear mathematical model

$$y_i = f(x_i, \Theta) + \epsilon_i, \quad i = 1, 2, ..., n$$

where

- y_i represents the observed data points,
- $f(x_i, \Theta)$ is a nonlinear function that depends on the parameter set Θ,
- x_i are the independent variables, and
- ϵ_i is an error term assumed to be normally distributed.

The goal of NLSM is to minimize the sum of squared residuals (RSS), defined as

$$Z(\Theta) = \sum_{i=1}^{n} [y_i - f(x_i, \Theta)]^2$$

where

- $Z(\Theta)$ is the objective function to be minimized,
- y_i are the actual observed values,
- $f(x_i, \Theta)$ represents the model's predicted values, and
- Θ is the set of unknown model parameters.

To evaluate model fitting, we use the relative error, given by

$$E_{\text{rel}} = \frac{\|I_{\text{est}}(t_i) - I_{\text{data}}(t_i)\|_2}{\|I_{\text{data}}(t_i)\|_2}. \tag{2}$$

This relative error, E_{rel}, quantifies the discrepancy between the estimated number of infected users $I_{\text{est}}(t_i)$ and the actual recorded infected users $I_{\text{data}}(t_i)$ at various time points t_i. The norm $\|\cdot\|_2$ represents the Euclidean distance, allowing for an objective measure of the overall deviation relative to actual data. A lower error value signifies a better model fit, affirming its accuracy in depicting the dynamics of toxicity spread.

We compared the error rates of the SEIQR model with and without dataset splitting (shown in Table III). The consistently low error rates across datasets when splitting the data into moderate and high toxicity levels, as compared to using the model without this division, underscore the effectiveness of the SEIQR model in capturing the dynamics of toxicity diffusion. The findings suggest that employing the SEIQR model and categorizing toxicity into moderate and high levels is reliable for understanding and predicting the spread of toxicity in various contexts. Hence, these analyses offer an answer to **RQ1**.

To answer **RQ2**, i.e., sensitivity analysis of the model parameters, we begin by identifying the parameter values that are most influential in determining social media toxicity spread (illustrated in Table IV). It is vital to discover numerous aspects that contribute to the toxicity spread and prevalence to decide the best technique for minimizing the number of affected users. To determine the dependence of each parameter on the SEIQR model, a sensitivity analysis of each parameter was performed using the Latin Hypercube Sampling-Partial Rank Correlation Coefficient (LHS-PRCC) method. The PRCC corresponds directly to the degree of statistical influence. A positive value indicates that an increase in this parameter leads to a positive influence on the SEIQR model. In contrast, a negative value indicates that an increase in this parameter leads to a negative influence on the SEIQR model. As shown in Figure 2, among the ϕ, β, ψ, μ, $\mathcal{A}$, and θ parameters, ϕ, β, ψ, $\mathcal{A}$, and θ have positive effects on

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-301-9

TABLE II. EXPLANATION OF THE MODEL PARAMETERS.

Parameter	Value	Source	Explanation
$\mathcal{A}$	[100]	fitted	recruitment rate of human
β	0.0006	fitted	effective contact rate
ψ	0.047	fitted	the rate at which exposed become infected
θ	0.020	fitted	the rate at which I transfer to quarantine class
η	0.04	fitted	the rate at which recovery becomes susceptible
γ	0.002	fitted	the rate at which Q transfer to recovery
ϕ	0.1	fitted	the rate at which I transfer to recovery
μ	0.1	fitted	the rate at which people exit autonomously

TABLE III. ERROR RATES FOR SEIQR MODEL WITH AND WITHOUT DATASET SPLITTING.

Dataset	Moderate	High	No split
F*Covid	0.0011	0.0021	0.081
F*Mask	0.021	0.049	0.073
Lockdown	0.032	0.045	0.061
5G	0.0029	0.0011	0.003
Brazil Anti	0.062	0.055	0.094
Brazil Pro	0.060	0.061	0.088
Peru	0.095	0.124	0.41

$\mathcal{R}_0$. In contrast, parameter μ has a negative effect on $\mathcal{R}_0$. The red line represents the p-value, and in the context of social media toxicity, $p < 0.05$ indicates that the results of the sensitivity analysis are statistically significant. Specifically, it means that there is less than a 5% probability that the observed effects (relationships between the parameters and the effective reproduction number $\mathcal{R}_0$) occurred by chance. These behaviors are significant in executing emergency management measures. Let,

$$C_p^{\mathcal{R}_0} = \frac{\partial R_0}{\partial p} \times \frac{p}{\mathcal{R}_0}$$

where, p is the parameter being studied for its sensitivity. Utilizing this index, one can deduce the sensitivity index corresponding to every parameter integrated into the expression for $\mathcal{R}_0$. We compute their derivatives as follows:

$$C_\beta^{\mathcal{R}_0} = \frac{\partial R_0}{\partial \beta} \times \frac{\beta}{\mathcal{R}_0},$$

such that,

$$\frac{dR_0}{d\psi} = \frac{A\beta(\mu+\phi+\theta-\psi)}{\mu(\mu+\psi)^2(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\beta} = \frac{A\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{dA} = \frac{\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\mu} = -\frac{A\beta\psi(\mu+\psi+\phi+\theta)}{\mu^2(\mu+\psi)(\mu+\phi+\theta)}$$

$$\frac{dR_0}{d\phi} = \frac{A\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)^2}$$

$$\frac{dR_0}{d\theta} = \frac{A\beta\psi}{\mu(\mu+\psi)(\mu+\phi+\theta)}$$

TABLE IV. SENSITIVITY INDEX OF EACH PARAMETER ON $\mathcal{R}_0$.

Parameters	Sensitivity Indices	Relationship
ψ	0.863213	+ve
β	0.902867	+ve
$\mathcal{A}$	0.8617	+ve
μ	-0.9167	-ve
ϕ	0.89574	-ve
θ	0.1913	+ve

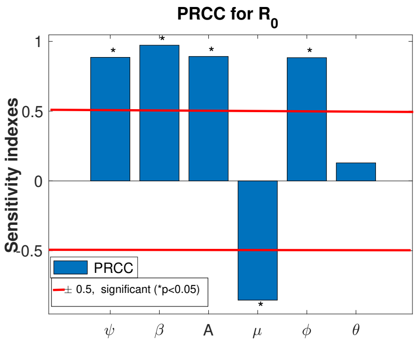

Figure 2. Sensitivity of $\mathcal{R}_0$ of the online toxicity contagion.

The positive sensitivity indices, associated with ψ, ζ, β, ϑ, σ, and ϵ, predominantly influence the frequency of online toxicity manifestations, with β having the highest positive impact index, indicating its crucial role in online toxicity spread (as illustrated in Figure 2). In contrast, negative sensitivity indices linked to parameters μ and ϕ attenuate the online toxicity spread. Remarkably, parameters such as the infection rate of exposed users (ϑ), recovery rates of users who hadn't been quarantined (σ), the rate at which infected users transfer to rehabilitation (ϵ), and the rate at which infected users are

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-301-9

banned from platform (ϕ) all curb the dissemination of online toxicity. This highlights the importance of improving measures to ban infected users and rehabilitation to control the spread of online toxicity effectively.

Building on the insights from the SEIQR model's numerical simulations and sensitivity analysis, several strategic interventions are recommended to control the spread of toxicity on social media platforms. These measures leverage fractional and fractal dynamics to enhance content moderation and behavioral management. **First**, we suggest adaptive moderation through behavioral memory and recurrence patterns. To effectively manage toxicity, platforms should implement temporary restrictions on accounts engaged in harmful behavior. These restrictions may include suspensions, limited posting capabilities, or increased scrutiny. This approach acknowledges the role of memory effects in user behavior, ensuring that past interactions influence future activities. By allowing monitored users to reform while preventing unchecked toxicity from gaining momentum, this method balances community safety with user rehabilitation. **Second**, we suggest algorithmic content demotion and network complexity reduction. Reducing the visibility of toxic content using algorithmic content demotion can limit its reach and engagement. By prioritizing non-toxic interactions and reducing amplification through engagement restrictions (e.g., limiting likes, shares, and comments on flagged content), the network's structure becomes less conducive to toxicity proliferation. This intervention effectively slows down the contagion effect. **Third**, we suggest cross-platform collaboration for unified toxicity control. Toxic content often spreads across multiple social media networks, requiring cross-platform collaboration for consistent moderation. A unified approach—incorporating shared databases, technological innovation, and standardized policies—can enhance intervention efficiency. Platforms can integrate collective quarantine mechanisms, ensuring that users flagged for toxicity on one platform face restrictions across multiple services, thereby mitigating repeated behavioral patterns. **Last**, we suggest structured quarantine and reintegration mechanisms. The SEIQR model suggests that quarantining toxic users is an effective strategy to limit further spread. Platforms should establish dedicated discussion spaces where quarantined users can engage in monitored interactions rather than simply migrating to another platform. This could include controlled forums, restricted access groups, or supervised comment sections, ensuring that users receive guidance toward positive behavioral reform.

By implementing these memory-aware and network-structured interventions, social media platforms can effectively mitigate the spread of toxicity, reduce harmful interactions, and foster a healthier digital environment.

V. Conclusion and Future Works

This study applied the SEIQR epidemiological model to analyze the propagation of toxicity on social media. By categorizing toxicity into moderate and high intensities, the model significantly improves prediction accuracy, as evidenced by lower error rates across diverse datasets. This granular classification aligns with real-world observations that toxicity manifests in varying degrees of severity, each requiring distinct intervention strategies. The sensitivity analysis further underscores the critical role of parameters such as the contact rate (β) and quarantine rate (θ) in modulating the basic reproduction number ($\mathcal{R}_0$), offering actionable insights for platform regulators.

Our key contributions include:

1) **Improved Predictive Accuracy:** Splitting toxicity into moderate and high levels reduces model error rates by up to 90% (e.g., Peru dataset error drops from 0.41 to 0.095), enabling precise resource allocation.
2) **Parameter Sensitivity:** The positive correlation of β, ψ, and $\mathcal{A}$ with $\mathcal{R}_0$ underscores the need to limit exposure to toxic content, while the negative impact of μ and θ validates quarantine policies.

However, the study has limitations. The reliance on threshold-based toxicity classification (e.g., Detoxify scores > 0.5) may oversimplify nuanced human communication. Additionally, the model assumes homogeneous mixing within compartments, which may not fully capture the fragmented nature of online communities. Future iterations could incorporate network-specific topology data to refine compartmental transitions.

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

[1] E. A. Vogels, *The state of online harassment*. Pew Research Center Washington, DC, 2021, vol. 13.

[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 2009, pp. 49–62.

[3] S. Shajari, M. Alassad, and N. Agarwal, "Characterizing suspicious commenter behaviors," in *Proceedings of the international conference on advances in social networks analysis and mining*, 2023, pp. 631–635.

[4] S. Shajari and N. Agarwal, "Safeguarding youtube discussions: A framework for detecting anomalous commenter and engagement behaviors," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 54, 2025.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

[5] S. Shajari and N. Agarwal, “Developing a network-centric approach for anomalous behavior detection on youtube,” *Social Network Analysis and Mining*, vol. 15, no. 1, p. 3, 2025.

[6] L. Dai, X. Liu, and Y. Chen, “Global dynamics of a fractional-order sis epidemic model with media coverage,” *Nonlinear Dynamics*, vol. 111, no. 20, pp. 19 513–19 526, 2023.

[7] R. Bernard, G. Bowsher, R. Sullivan, and F. Gibson-Fall, “Disinformation and epidemics: Anticipating the next phase of biowarfare,” *Health Security*, vol. 19, no. 1, pp. 3–12, 2021, PMID: 33090030. DOI: 10.1089/hs.2020.0038. eprint: https://doi.org/10.1089/hs.2020.0038. [Online]. Available: https://doi.org/10.1089/hs.2020.0038.

[8] P. Petratos and A. Faccia, “Fake news, misinformation, disinformation and supply chain risks and disruptions: Risk management and resilience using blockchain,” *Annals of Operations Research*, vol. 327, no. 2, pp. 735–762, 2023.

[9] K. DiCicco, N. B. Noor, N. Yousefi, M. Maleki, and N. Agarwal, “Toxicity and networks of covid-19 discourse communities: A tale of two social media platforms,” in *Proceedings of ROMCIR 2023: The 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2023: the 45th European Conference on Information Retrieval, April 2-6, 2023, Dublin, Ireland*, pp. 1–13.

[10] N. Yousefi and N. Agarwal, “Study the influence of toxicity intensity on its propagation using epidemiological models,” in *Proceedings of the 30th Americas Conference on Information Systems (AMCIS)*, 2024, pp. 2401–2410.

[11] D. DeMarsico, N. Bounoua, R. Miglin, and N. Sadeh, “Aggression in the digital era: Assessing the validity of the cyber motivations for aggression and deviance scale,” *Assessment*, vol. 29, no. 4, pp. 764–781, 2022.

[12] N. B. Noor, N. Yousefi, B. Spann, and N. Agarwal, “Comparing toxicity across social media platforms for covid-19 discourse,” in *The Ninth International Conference on Human and Social Analytics (HUSO)*, 2023, pp. 21–26.

[13] N. Yousefi, N. B. Noor, B. Spann, and N. Agarwal, “Towards developing a measure to assess contagiousness of toxic tweets,” in *Proceedings of the international workshop on combating health misinformation for social wellbeing*, 2023, pp. 43–47.

[14] N. Yousefi, N. B. Noor, B. Spann, and N. Agarwal, “Examining toxicity’s impact on reddit conversations,” in *International conference on complex networks and their applications*, Springer, 2023, pp. 401–411.

[15] S. Dagtas, N. Agarwal, and N. Yousefi, “Modeling toxicity propagation on reddit using epidemiology,” in *International Conference on Complex Networks and Their Applications*, Springer, 2024, pp. 113–124.

[16] T. C. Falade, N. Yousefi, and N. Agarwal, “Toxicity prediction in reddit,” in *In Proceedings of the 30th Americas Conference on Information Systems (AMCIS)*, 2024, pp. 2835–2844.

[17] T. Li, S. Wang, and B. Li, “Research on suppression strategy of social network information based on effective isolation,” *Procedia computer science*, vol. 131, pp. 131–138, 2018.

[18] M. Maleki and N. Agarwal, “A comparative evaluation of the sir and seiz epidemiological models to describe the diffusion characteristics of covid-19 polarizing viewpoints on online social networks,” in *In Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*, 2025, pp. 2483–2492.

[19] M. Maleki, M. Arani, E. Mead, J. Kready, and N. Agarwal, “Applying an epidemiological model to evaluate the propagation of toxicity related to covid-19 on twitter,” in *In Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, 2022, pp. 3275–3284.

[20] R. Das and W. Ahmed, “Rethinking fake news: Disinformation and ideology during the time of covid-19 global pandemic,” *IIM Kozhikode Society & Management Review*, vol. 11, no. 1, pp. 146–159, 2022.

[21] G. A. Ngwa and M. I. Teboh-Ewungkem, “A mathematical model with quarantine states for the dynamics of ebola virus disease in human populations,” *Computational and mathematical methods in medicine*, vol. 2016, no. 1, pp. 1–29, 2016.

[22] E. Addai, N. Yousefi, and N. Agarwal, “Seiqr: An epidemiological model to contain the spread of toxicity using memory-index,” in *Fifth International Workshop on Cyber Social Threats, International Conference on Web and Social Media*, 2024, pp. 11–22.

[23] L. Hanu, “Unitary team. detoxify,” *Github: https://github.com/unitaryai/detoxify [last accessed: September 3, 2025]*, 2020.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-301-9