Applying an Epidemiological Model to Evaluate the Propagation of Toxicity related to COVID-19 on Twitter

Maryam Maleki
COSMOS Research Center, UA- Little Rock
mmaleki@ualr.edu

Mohammad Arani
Systems engineering, UA-Little Rock
mxarani@ualr.edu

Esther Mead
COSMOS Research Center, UA- Little Rock
elmead@ualr.edu

Joseph Kready
COSMOS Research Center, UA- Little Rock
jkready@ualr.edu

Nitin Agarwal
COSMOS Research Center, UA- Little Rock
nxagarwal@ualr.edu

Abstract

The prevalence of social media has increased the propagation of toxic behavior among users. Toxicity can have detrimental effects on users' emotion and insight and disrupt beneficial discourse. Evaluating the propagation of toxic content on social networks such as Twitter can provide the opportunity to understand the characteristics of this harmful phenomena. Identifying a mathematical model that can describe the propagation of toxic content on social networks is a valuable approach to this evaluation. In this paper, we utilized the SEIZ (Susceptible, Exposed, Infected, Skeptic) epidemiological model to find a mathematical model for the propagation of toxic content related to COVID-19 topics on Twitter. We collected Twitter data based on specific hashtags related to different COVID-19 topics such as covid, mask, vaccine, and lockdown. The findings demonstrate that the SEIZ model can properly model the propagation of toxicity on a social network with relatively low error. Determining an efficient mathematical model can increase the understanding of the dynamics of the propagation of toxicity on a social network such as Twitter. This understanding can help researchers and policymakers to develop methods to limit the propagation of toxic content on social networks.

1. Introduction

There is evidence of a growing population of users on social media platforms that post and share content that is considered "toxic" in that it contains profanity, insults, sexual themes, etc. These toxic users can disrupt the principles of a social media platform and can cause harmful effects on other users' emotions and opinions. In this study, we adopt the operational definition of toxicity from previous literature: "the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion" [1]–[3]. The propagation of toxicity may have significant effects on different aspects of people's lives. However, spreading toxicity and inappropriate insight about healthcare subjects can be more harmful and can pose a serious threat to people's health.

This study is motivated by how toxicity could influence people's behaviors on social media, specifically relative to public health issues. We attempted to apply a mathematical model to explain how toxicity propagates on Twitter, particularly for COVID-19 discourse. We collected four different datasets containing COVID-19 hashtags during the entirety of 2020. We were motivated to apply a particular epidemiological model to determine the diffusion trends of toxicity on Twitter. The basis of epidemiological models involves dividing the population into different compartments that each represent the state of an individual involved in the considered social network.

In this research, we have applied the SEIZ (Susceptible, Exposed, Infected, and Skeptic) model, which is a strong model for the propagation of information and ideas compared to the other epidemiological models [4]. Unlike traditional epidemiological models (e.g., SIS and SIR), the SEIZ model has an additional compartment for Exposed (E) individuals, who do not react immediately to the information they receive on social media and need some time to become *infected* by the information. The Infected group is defined as users whose posts (or tweets in the case of Twitter) contain toxicity. Moreover, this model has a Skeptic (Z) compartment, which contains users who have seen the post, but are indifferent to the information and decided not to engage in any response to it. To the best of our knowledge, there has been no prior study which has empirically applied an epidemiological model to the propagation of toxicity on

URI: https://hdl.handle.net/10125/79735
978-0-9981331-5-7

Twitter. The main goal of this study is to find a mathematical model that explains the propagation of toxicity regarding COVID-19 issues on Twitter. A robust mathematical model for the propagation of toxicity on online social networks can enable researchers to evaluate the number of users in any compartment at any time. In these models, however, the Infected compartment is of primary importance since it is composed of the users who actively spread the toxicity.

The remainder of this paper is organized as follows. Section 2 presents the related work that has been done regarding the concept underlying the spread of emotion, the existence of toxicity on social media, and efforts regarding the application of epidemiological modeling to online social networks. In section 3, the methodologies for data collection and analysis are explained. We also describe the basics of two of the traditional epidemiological models (SIS, SIR) and then these models are compared to the SEIZ model, which is utilized in this study. In section 4, we discuss the overarching themes and impact of this research. Finally, section 5 concludes the paper with ideas for future work.

2. Literature Review

People are struggling emotionally during the current Coronavirus pandemic, especially on social media platforms, and they often express and share their feelings. A large body of evidence suggests that toxic expressions are almost always accompanied by negative emotions [5]. As there are more works on modeling the spread of emotion than the spread of toxicity on social media, we reviewed these works to get inspiration to understand how to model the spread of toxicity. To study the issue on social media from a literature standpoint, reviews have focused on emotion, toxicity, and epidemiological models.

2.1. Spread of Emotion

Recent studies suggest that similar to in-person communications, human emotions also disseminate through conscious and unconscious pathways [6]. Emotional states such as joy, sadness, trust, disgust, fear, anger, surprise, and anticipation could be contagious in online social media discourse and the influence could be daunting and intimidating from the network's circles to users. However, the underlying mechanisms of emotional contagion in social media are rarely investigated. Kramer et al. studied controlled news feeds provided to users resulting in experimental evidence for emotional contagion via the Facebook network [6]. In another study that divided individuals into two classes of highly and scarcely susceptible to emotional contagion, Ferrara et al. [7] showed that there is a linear relationship between the average positive and negative emotions of the post that users are exposed to, and that of their response they have to that stimuli post. Additionally, the scarcely susceptible users were more likely to espouse negative emotions. Kwon and Gruzd [8] studied the spread of blasphemy by two mechanisms known as mimicry and social interaction effect on YouTube in which public swearing starts a chain of interpersonal swearing. The study is based upon mixed-effect logistic regression models and data were composed of offensive comments in reply to the 2016 U.S. presidential campaign. The most recent study by Fan and et al. [9] examined the spread of angry tweets on Weibo, a Chinese microblogging site similar to Facebook. Employing a diffusion model illustrated that weaker social network ties accelerated propagation of anger with respect to the metrics of velocity and coverage. The authors also found that strangers have a greater tendency to broadcast rage rather than joy.

2.2. Toxicity on Social Media

Cheng et al. [10] demonstrated that toxic users become worse over time as measured by the content they post. In another study, Cheng et al. [3] concluded that given the right condition, anyone can exhibit toxic tendencies. A comprehensive examination of various forms of online toxicity was conducted by Warner et al. [11]. Researchers have proposed ways to identify and mitigate hate speech (toxicity) in online communities [12], [13]. Wulczyn et al. [13] applied machine learning techniques including linear regression and multilayer perceptron in an attempt to identify personal attacks at scale, concluding that the problem remains surprisingly difficult. To aid in the identification endeavor, Davidson et al. [14] presented a dataset with three kinds of comments: hate speech, offensive but non hateful speech, and neither. In another study, five different forms of toxicity between the comments posted on "pro- and anti-NATO" channels on YouTube were evaluated [1]. They authors used the "YouTube Data API" and the YouTubeTracker tool [16] to collect a large dataset of YouTube comments for analysis. They then assigned a toxicity score to each comment using "Google's Perspective API". Their analysis demonstrated that comments on pro-NATO channels were less toxic than those on the anti-NATO channels. In another paper, Obadimu et al. propoosed an epidemiological model to evaluate the spread of toxicity on YouTube [15]. Termed the STRS (Susceptible, Toxic, Recovered, Susceptible) model, the authors proposed that there is a similarity between the propagation of toxicity on YouTube and the spread of a disease in a population. Their paper adopted a theoretical approach, wherein no

real-world data was used to evaluate the STRS model. We conducted a subsequent comparative analysis, however, which showed that the SEIZ model outperformed the STRS model. Due to the page limitation of this current work, our evaluation between the STRS and the SEIZ models has not been included in this current study. However, we do plan to include that evaluation along with the evaluation of other epidemiological models in our extended version of this work.

2.3. Epidemiological Modeling in Social Networks

Applying a mathematical model to evaluate the spread of ideas on an Online Social Network (OSN) can provide us with the opportunity to acquire effective information toward its propagation. As a result, we can set the stage for useful approaches and policies to control this propagation [17]. The basic framework for all epidemiological models involves dividing the population into different compartments. The primary epidemic model is the SI (Susceptible- Infected) model, which partitions the total population into Susceptible and Infected compartments based on disease state. In this model, the Infected compartment involves individuals who are already carrying the infection, while the Susceptible compartment consists of people who have not yet acquired the infection but are at risk of contracting the infection from Infected individuals [18]. Moreover, people who are infected may be transferred to the Susceptible compartment again, which is part of the SIS (Susceptible-Infected-Susceptible) model [15]. The SIR (Susceptible-Infected-Recovered) model is another epidemiological model frequently used in different studies. This model includes the Recovered compartment, which involves individuals who develop immunity to the infection [19]–[21].

The similarity between the propagation of a disease and a rumor in mathematical terms was first studied by Daley and Kendall [22], [23]. Over the years, different epidemiological models derived from the SIR model were applied to evaluate the propagation of information and rumors in a population [24]. Abdullah et al. (2011) applied the SIR model to study the spread of news on Twitter. Their findings acknowledged their hypothesis about the similarity between the propagation of disease and the spread of information on Twitter [20]. In another study, Jin et al. (2013) used an epidemiological model to evaluate the propagation of news and rumors on Twitter. The authors applied the SEIZ model to evaluate the diffusion trends of four news items and four rumors on Twitter. Their model includes a Skeptic (Z) compartment, which consists of users who know about the story but decided not to spread it. It also includes Exposed (E) individuals, which are users who know about the news but needed some time to decide whether to spread it [17].

3. Methodology

In this section, methods used for data collection are described. We then discuss the application of epidemiological models. We then provide a detailed description of the SEIZ model, which was ultimately used as the model for our datasets.

3.1. Data Collection and Processing

We used Twitter Academic APIs to collect tweets related to COVID-19 for the entire year of 2020. We collected data for different hashtags that could best cover a broad range of topics related to COVID-19. These topics included the following: lockdown, mask, and vaccine. These hashtags were chosen after doing a qualitative analysis on Twitter to find frequent and commonly used hashtags during the pandemic. Our dataset contains original tweets as well as retweets and replies. There was no language restriction as the toxicity computation also works for non-English languages. The list of identified hashtags and their respective number of tweets is shown in Table 1. Table 1 also includes the error metric relative to our experiments, which we discuss in more detail in section 4.3.

Table 1. Tweet counts for hashtags related to COVID-19 topics

Hashtag	No. of Tweets	First tweet	Last tweet	Error
#f*ckmasks	3,456	2020-02-07	2020-12-31	0.063
#f*ckvaccine	2,735	2020-01-03	2020-12-31	0.049
#f*cklockdown	1,995	2020-03-14	2020-12-31	0.112
#f*ckcovid	45,569	2020-02-12	2020-12-31	0.058

To provide contextual insights, Table 2 provides some examples of tweets containing the identified hashtags and their respective toxicity scores.

Table 2. Example tweet from each dataset

Dataset	Post	Text-toxicity
#f*cklockdown	If we get locked down for 2 weeks Boris Johnson can s*ck my left ball if he thinks I'm going to stay in my house for that long #f*ckthelockdown	0.84
#f*ckmask	F*ck mask in public. i wear glasses, that sh*t doesn't work for me. Ya know we be fogged up with that sh*t. Cloth doesn't cover u from sh*t. get a gas mask if u scared #f*ckmask	0.99
#f*ckvaccine	China unleashes a virus that bankrupts most of the world and now governments want to force experimental vaccines on its citizens. Is any country going to hold these f*ckers responsible for this? #fuckthevaccine	0.95
#f*ckcovid	I'm ashamed to call myself and American, we really failed to contain this sh*tty virus 👇 #F*ckAmerica #F*ckCovid	0.87

3.2. Model

To evaluate the propagation of toxicity on social networks, and specifically on Twitter, we used an epidemiological model, which divides the population into different compartments.

3.2.1. SIS Model

The SIS model is one of the preliminary epidemiological models which divides the population into two parts: Susceptible (S) and Infected (I) (Figure1). Since there is no accounting for immunity against the infection in the SIS model, the Infected individual returns to the Susceptible compartment. To adapt this model to the idea of the spread of the toxicity on Twitter, we used a new definition for these groups. A user is Infected if they post a tweet using a hashtag identified in our qualitative analysis as one with the potential of propagating toxicity, and Susceptible if they have not yet posted tweets using the mentioned hashtag. When a Susceptible user contacts an Infected user, the user will become Infected and will post a tweet using the hashtag [16].

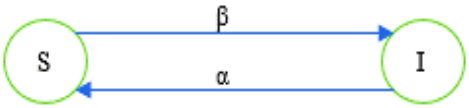

Figure 1. SIS model

3.2.2. SIR Model

Another model, which is more often used in different studies and is more practical than the SIS model, is the SIR model. This model divides the population into three different parts: Susceptible (S), Infected (I), and Recovered (R) (Figure 2). In this model, Infected people consist of people who have the infection and can spread it to others. Susceptible people are individuals who are at risk of becoming infected. The Recovered people are those who are immune from the infection or have died from the infection; consequently, they cannot cause another person to become infected [19].

To adjust this model to the spread of toxicity on Twitter, we allocated new definitions to these terms. A user is Infected if they post a tweet using a hashtag identified in our qualitative analysis as one with the potential of propagating toxicity and they are Susceptible if they follow the Infected person and have not yet posted a tweet containing the specific hashtag themselves. They are Recovered if they have not subsequently posted tweets containing the specific hashtag within a certain time frame.

Figure 2. SIR model

3.2.3. SEIZ Model

One important restriction of the SIS and SIR traditional epidemiological models is that when a Susceptible individual encounters an Infected user, there is just one possible action, which is that the user can becomes Infected. However, this assumption does not apply properly to the propagation of toxicity, specifically on social media. Users may have different mindsets when they are exposed to toxicity on social media. When people are exposed to toxicity on social media, they may be convinced to further propagate that toxicity after some consideration. This decision could be immediate for some users, while for others it may take some amount of time.

Moreover, it is possible that some users are never affected by this toxicity and do not show any reaction to tweets that contain toxicity. These scenarios are possible

but are not covered by the basic SIS and SIR epidemiological models. In the context of analyzing the propagation of toxicity on Twitter, the different compartments of the SEIZ model (Figure 3) are outlined below.

- Infected (I) relates to users who have posted tweets using a hashtag identified in our qualitative analysis as one with the potential of propagating toxicity.
- Susceptible (S) represents users who follow the Infected individuals and are at the risk of getting infected via the contact.
- Exposed (E) represents the users who have been Exposed to the tweets containing an identified hashtag and had a delay of time before posting an additional tweet using the specific hashtag.
- Skeptic (Z) refers to individuals who have encountered the toxicity via a tweet but decide not to tweet and use the hashtag [17].

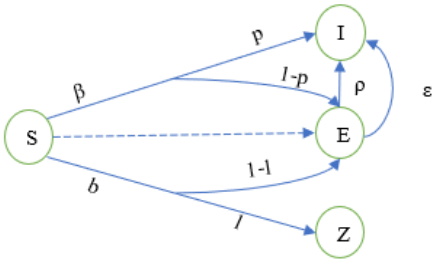

Figure 3. SEIZ model

The following system of Ordinary Differential Equations (ODE) represents the SEIZ model [17].

$$\frac{dS}{dt} = -\beta S\frac{I}{N} - bS\frac{Z}{N} \quad (1)$$

$$\frac{dE}{dt} = (1-p)\,\beta S\frac{I}{N} + (1-l)bS\frac{Z}{N} - \rho E\frac{I}{N} - \varepsilon E \quad (2)$$

$$\frac{dI}{dt} = p\beta S\frac{I}{N} + \rho E\frac{I}{N} + \varepsilon E \quad (3)$$

$$\frac{dZ}{dt} = lbS\frac{Z}{N} \quad (4)$$

For the above-mentioned ODEs, the parameters are defined in Table 3.

Table 3. Parameters of the SEIZ model

Parameter	Definition
β	Contact rate between S and I.
b	Contact rate between S and Z.
ρ	Contact rate between E and I.
p	Probability of S to I given contact with I.
1-p	Probability of S to E given contact with I.
ε	Transition rate of E to I (Incubation rate).
l	Probability of S to Z given contact with Z.
1-l	Probability of S to E given contact with Z.

When a Susceptible (S) (the user who follows an Infected (I) user) comes into contact with the Infected person (I) with β rate, they can immediately decide to share the tweet with p probability, or that user may need some time to think about it and move to the Exposed (E) compartment with (1-p) probability.

In addition, a Susceptible may come into contact with a Skeptic (Z) (a user who saw the tweet containing toxicity but decided not to tweet about it) with the rate b. This contact can lead to two different scenarios. The first possibility is that it can lead to turning the user into another Skeptic with the probability of l. This means that the user chose not to tweet about it or to not tweet using the specific hashtag. The second possibility is that the contact may result in the unintentional outcome of leading the user into the Exposed (E) compartment with the probability (1-l). Transferring users from the Exposed state to the Infected state can happen from two different scenarios. The first possibility is that the Exposed (E) (user who has heard about the hashtag but needs some time before tweeting about it and sharing the hashtag) may have more contact with Infected users with a contact rate ρ and because of this further contact they will become Infected. The second possibility is that the Exposed (E) user can move to the Infected compartment not because of contacting Infected users, but because of self-adoption with rate ε.

4. Analysis and Results

This section presents the research findings in three different parts. First, a preliminary analysis evaluates the frequency of the usage of the identified hashtags over time. Second, we discuss our toxicity analysis. Finally, the SEIZ model was applied to fit our different datasets to the Infected (I) compartment of the model.

4.1. Frequency of Tweets

In this section we analyze the frequency and cumulative sum of the tweets for different datasets. Again, these datasets were created based on hashtags that were identified in our qualitative analysis as ones with the potential of propagating toxicity. Due to space limitations, frequency and cumulative sum figures for all datasets are available upon request. As we can see in the Figure 4, the first hashtags related to masks in 2020 were spread in early February, but with negligible frequency until around the beginning of May. After May, however, mask-related hashtag usage started to increase with two big spikes in early July and early October. The frequency of tweets for other hashtags are shown in figures 5, 6, and 7. The trends of propagation of tweets and their peaks can be seen in the mentioned figures.

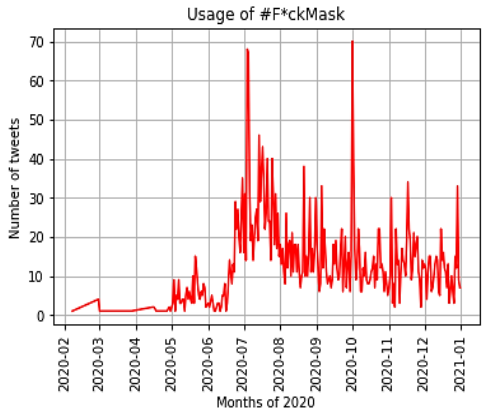

Figure 4. Frequency of tweets for hashtag related to mask

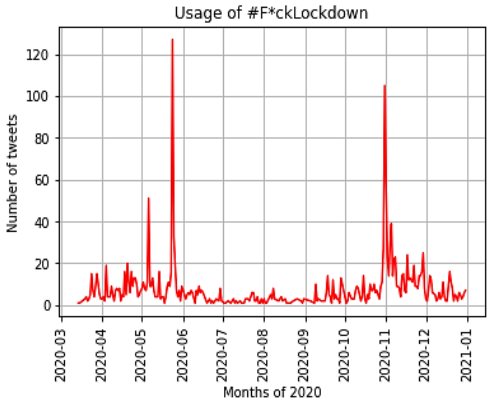

Figure 5. Frequency of tweets for hashtag related to lockdown

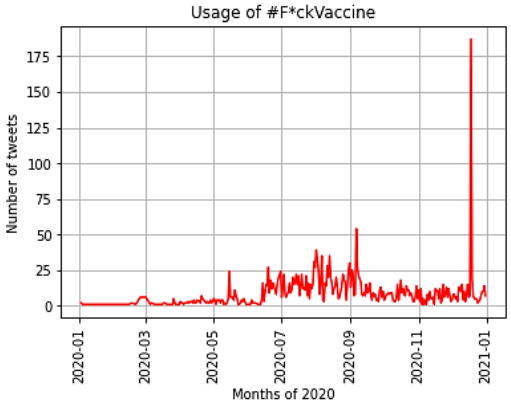

Figure 6. Frequency of tweets for hashtag related to vaccine

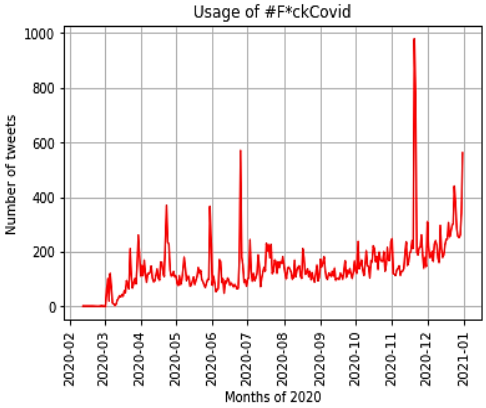

Figure 7. Frequency of tweets for hashtag related to covid

The spread of toxicity related to covid started in early February as well, with significant spikes in late June and late November (Figure 7). Figure 8 reveals that the cumulative sum of the tweeting activity increased at a relatively consistent rate throughout the year.

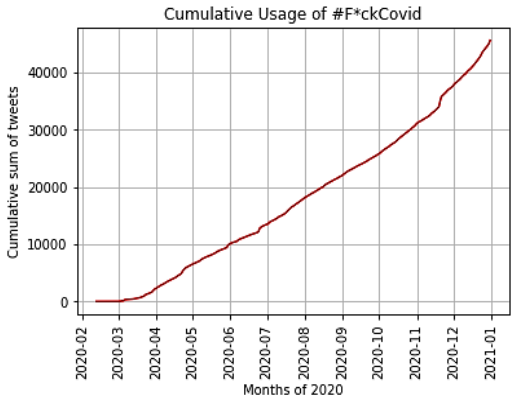

Figure 8. Cumulative sum of tweets for hashtag related to covid

When viewed on a cumulative scale, we can see that the overall posting frequency of hashtags identified as having the potential of propagating toxicity regarding masks is more prolific than for those regarding lockdown and vaccine (Figure 9). Interestingly, however, there are two considerable spikes within the lockdown dataset (Figure 5), late May and late October. Of further note, an s-shaped trend curve can be seen within the cumulative spread for both masks and vaccines (Figure 9). Within the cumulative spread related to lockdown, however, there are two apparent s-shaped trend curves (Figure 9). These s-curves are indicative of the adoption of the use of these hashtags in terms of tweeting behavior. We explore these concepts further in future works.

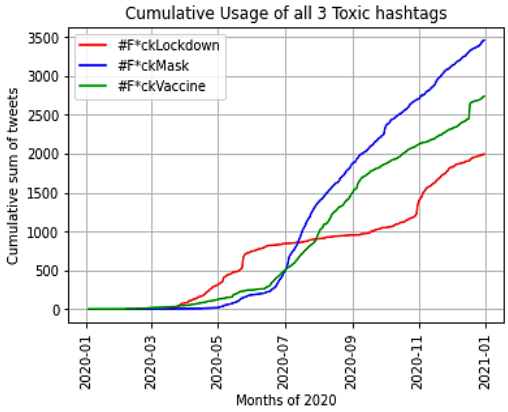

Figure 9. Cumulative sum of tweets for hashtags related to lockdown, mask, and vaccine

4.2. Computing Toxic Scores

Although there is no doubt that the collected hashtags related to these different topics of COVID-19 are toxic by nature, assigning a toxicity score for each post in our datasets is imperative. To do this, we used the Unbiased Detoxify Model which is a model from the 2019 "Jigsaw Unintended Bias in Toxicity Classification" challenge [25]. This particular toxicity challenge model outputs seven toxicity scores: a) "text_toxicity", which is the overall score for the text input (in this case, a tweet), b) "severe_toxicity", to identify the probability that a text input will be considered as severely toxic, and then five additional categories to identify the type of toxicity:

c) "obscene:
d) "threat"
e) "insult"
f) "identity_attack"
g) "sexual_explicit"

This model returns a probability score between 0 and 1, where higher values indicate a higher probability of the toxicity label being applied to the text input. We analyzed the scores for each of these toxicity components for each tweet in our four datasets (F*ckmask, F*cklockdown, F*ckvaccine, and F*ckcovid). The average overall toxicity scores for the datasets are 0.61, 0.51, 0.55, 0.42 respectively. Also, the average toxicity score for the combined datasets is 0.52. These scores confirm that our collected datasets contain toxicity, wherein "toxic" content is defined as a unit of text input being assigned a toxicity score of 0.5 or greater [26].

Table 3 shows the percentage of posts that are toxic within each dataset. The "severe_toxicity" component was not included in our further analysis due to its low count (proportion) within each of our datasets. Additionally, a very low proportion of the tweets in our datasets fell within the "threat", "identity_attack", and "sexual_explicit" categories. Therefore, our subsequent analysis only includes that for the "text_toxicity", "obscene" and "insult" toxicity categories. The highest average score for the obscene category was for posts related to vaccine, while the highest average score for text_toxicity and insult was for posts related to mask (Figure 10).

Table 3. Percentage of Toxic posts (toxicity score of 0.5 or greater) for different categories of toxicity

Toxicity Categories	Mask	Lockdown	Vaccine	Covid	Average
Text toxicity	64.76	54.69	57.22	38.84	53.88
Obscene	59.35	99.9	55.98	37.65	63.22
Threat	0.06	0.11	0.04	0.08	0.0725
Insult	23.93	13.99	7.35	9.32	13.6475
Identity attack	0.09	0.06	0	0.09	0.06
Sexual	1.25	1.41	0.26	1.2	1.03

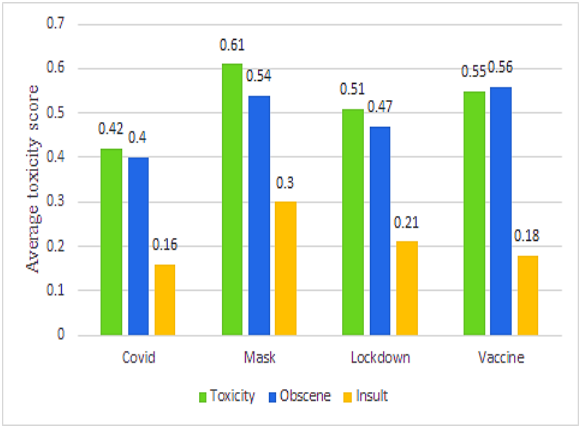

Figure 10. Average score for three different categories of toxicity for each dataset

Figure 11 shows the percentage of toxic tweets within the text_toxicity, obscene, and insult toxicity categories for each dataset. One of the interesting findings is that 99.9 percent of the posts related to lockdown have a score greater than 0.5 for obscene which may indicate a high level of anger or otherwise negative sentiment from users regarding the topic of mandatory lockdowns due to COVID-19.

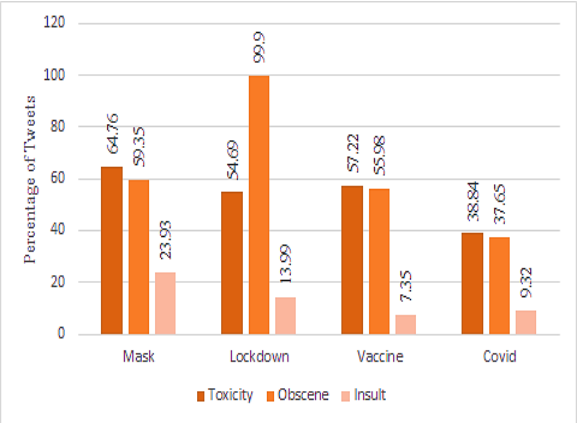

Figure 11. Percentage of toxic tweets (toxicity score of 0.5 greater) by toxicity category for each dataset

Figure 12 shows the evolution of the average toxicity score (the text_toxicity component of the model) over the different months of 2020 for each dataset. The toxicity scores for the tweets related to vaccine have the most fluctuation during different months. Also, it can be inferred from the figure that the tweets related to mask have the highest toxicity score in some months (April, August, September, November, December). The trend of toxicity for the covid dataset is more stable compared to the others, and it has a lower toxicity score than each of the others for about five out of 11 months within the year 2020.

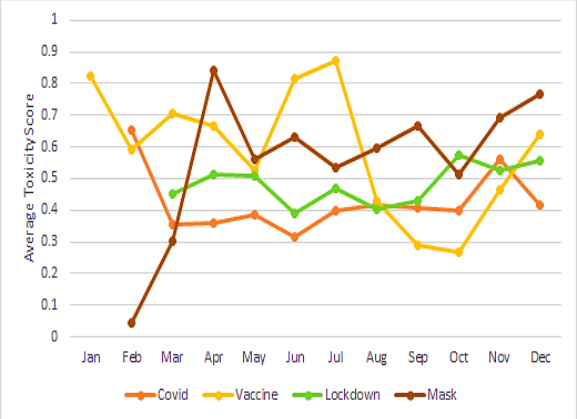

Figure 12. Evolution of average toxicity score (text_toxicity) for the year 2020

4.3. Fitting datasets to Infected (I) component of the SEIZ model

We fit the number of Infected people (those users who used the hashtags in each experiment) in each 24-hour time interval as the Infected (I) compartment in the SEIZ model by using MATLAB. We used the lsqnonlin [1] function, which is a nonlinear least square curve fitting function to fit our model to each of the four datasets. To solve the ODEs, we used ode45. Results were obtained from a laptop with Intel Core i5 CPU and 12 GB of RAM.

For every dataset there are a set of optimal parameters which can minimize the error between the actual number of tweets in the Infected compartment (i.e., users of hashtags) and the estimated number of users in the Infected compartment, |I(t) - tweets(t)|. Parameter tables for all datasets are available upon request. Model fit results for hashtags were graphed in Figures 13 through 16. The blue dots are the actual tweets while the red line is the Infected (I) compartment of the SEIZ model. While the end point for tweets was the same for each of the hashtags, the starting times were different (Table 1).

The "error" column in Table 1 displays the difference between the actual number of tweets containing the hashtag and the Infected compartment predicted by the SEIZ model, reflecting the relative error in 2-norm [17]. When comparing the results for each of the four experiments, the lowest error was obtained in vaccine (0.049); whereas, the highest error was obtained in lockdown (0.112). However, based on the relatively low error obtained in our experimental results, we conclude that the SEIZ model can be appropriate for modeling the spread of toxicity as based on different hashtags related to COVID-19.

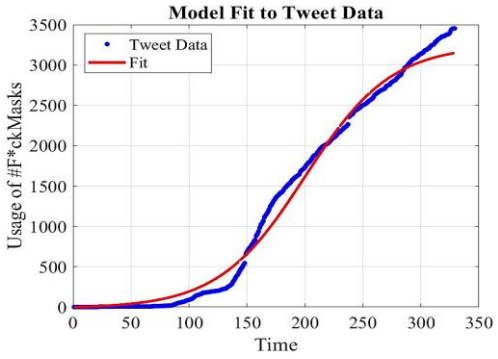

Figure 13. SEIZ model fit for the hashtag #F*ckmasks

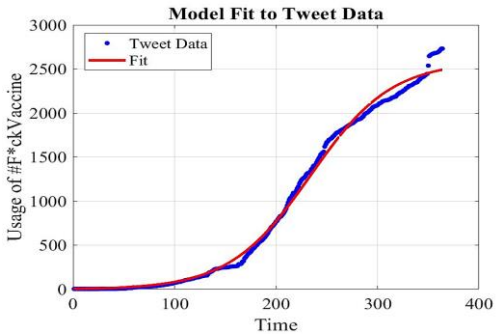

Figure 14. SEIZ model fit for hashtag #F*ckvaccine

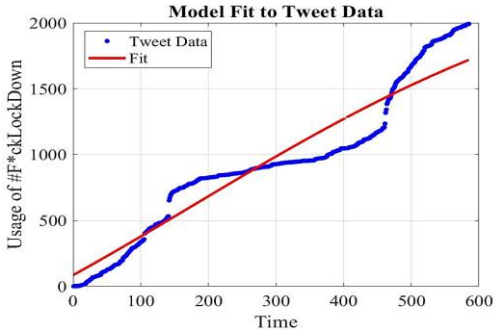

Figure 15. SEIZ model fit for the hashtag #F*cklockdown

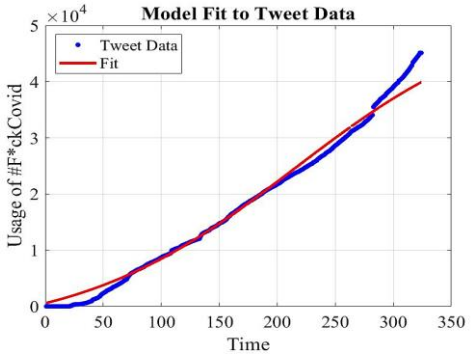

Figure 16. SEIZ model fit for the hashtag #F*ckcovid

The purpose of this study was to determine whether there are similarities among the spread of epidemics in the real world and toxic expressions on social media. The latest epidemiology model employed in our study is the SEIZ model. To summarize, the key findings of this study include:

- Based on the error metrics calculated for the four datasets used in our experiments, the SEIZ model predicts the Infected compartment quite well in terms of the spread of toxicity on Twitter.
- The trends that are observed in the extracted datasets could be divided into three general categories, (a) s-shaped curves such as in Figures 13 and 14, (b) tangent-shaped curves such as that in Figure 15, and (c) straight (or close to straight) lines such as that in Figure 16.
- Although the overall performance of the SEIZ model was promising and satisfactory to make predictions in the case of s-shaped curves and in cases of straight (or close to straight) lines, fitting tangent-shaped curves was challenging, producing the highest level of error (the case in our lockdown dataset as shown in Figure 15).
- This work reveals a limitation of the SEIZ model, which is its apparent inability to model the propagation of toxicity on a social network, specifically Twitter, in cases of tangent-shaped toxicity dissemination trends.
- According to Tables 1 and 3 and Figure 11, three out of our four toxic datasets with toxicity scores of 0.5 or greater and representing a proportion of content ranging from 38.84% to 64.76% were estimated by the SEIZ model with the lowest possible error. These three datasets are covid, vaccine, and mask. However, the lockdown dataset presents a higher error of fit due to its tangent-shaped trend.

5. Conclusions and Future Work

In this study, we demonstrated how the propagation of hashtags related to COVID-19 on Twitter can be modeled by applying the SEIZ epidemiological model. We applied the I compartment of the SEIZ model to four different datasets containing hashtags identified in our qualitative analysis as ones with the potential of propagating toxicity for the subjects of mask, vaccine, lockdown, and covid. While the modeling error was relatively high for one dataset (lockdown: 0.112), it was relatively low for the remaining datasets (mask: 0.063; vaccine: 0.049; covid: 0.058). Such findings illustrate the strength of the SEIZ model to model the spread of toxicity on social media. Using mathematical models to study the spread of toxicity on social media, especially Twitter, can provide an opportunity to predict its trend. This can help policymakers to develop suitable strategies for controlling and preventing the spread of toxicity.

In future work, we plan to apply the SEIZ model to datasets collected from other social media platforms, such as Reddit, Facebook, YouTube, and Instagram. In addition, we plan to fit the datasets to the other compartments of the SEIZ model, such as the Skeptic (Z) compartment. As a result, we hope to find a way to transfer more users from the Susceptible compartment to the Skeptics compartment, which refers to users who decide not to respond or engage with toxic discourse, thereby preventing further spread of toxicity *infections*. Additionally, we will apply other epidemiological models on these datasets, compare them to previously published results, and determine which epidemiological models have the best performance. Future research also includes the application of epidemiological models to the study the spread of toxicity in various other domains, such as politics, healthcare, and religion.

Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-

15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). 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.

7. References

[1] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying toxicity within youtube video comment," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11549 LNCS, pp. 214–223, 2019.

[2] M. Märtens, S. Shen, A. Iosup, and F. Kuipers, "Toxicity detection in multiplayer online games," in *2015 International Workshop on Network and Systems Support for Games (NetGames)*, 2015, pp. 1–6.

[3] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone Can Become a Troll," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 1217–1230.

[4] L. M. A. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez, "The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models," *Phys. A Stat. Mech. its Appl.*, vol. 364, pp. 513–536, 2006.

[5] S. Steinert, "Corona and value change. The role of social media and emotional contagion," *Ethics Inf. Technol.*, Jul. 2020.

[6] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Natl. Acad. Sci.*, vol. 111, no. 24, pp. 8788–8790, Jun. 2014.

[7] E. Ferrara and Z. Yang, "Measuring emotional contagion in social media," *PLoS One*, vol. 10, no. 11, p. e0142390, 2015.

[8] K. H. Kwon and A. Gruzd, "Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos," *Internet Res.*, vol. 27, no. 4, pp. 991–1010, Aug. 2017.

[9] R. Fan, K. Xu, and J. Zhao, "Weak ties strengthen anger contagion in social media," May 2020.

[10] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[11] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*, 2012, pp. 19–26.

[12] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 71–80.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1391–1399.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Eleventh international aaai conference on web and social media*, 2017.

[15] A. Obadimu, E. Mead, M. Maleki, and N. Agarwal, "Developing an Epidemiological Model to Study Spread of Toxicity on YouTube," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2020, pp. 266–276.

[16] "YouTubeTracker – COSMOS." .

[17] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th workshop on social network mining and analysis*, 2013, pp. 1–9.

[18] K. M. A. Kabir, K. Kuga, and J. Tanimoto, "Analysis of SIR epidemic model with information spreading of awareness," *Chaos, Solitons & Fractals*, vol. 119, pp. 118–125, 2019.

[19] V. Colizza and A. Vespignani, "Invasion threshold in heterogeneous metapopulation networks," *Phys. Rev. Lett.*, vol. 99, no. 14, p. 148701, 2007.

[20] S. Abdullah and X. Wu, "An epidemic model for news spreading on twitter," in *2011 IEEE 23rd international conference on tools with artificial intelligence*, 2011, pp. 163–169.

[21] H. Wang *et al.*, "A variant epidemic propogation model suitable for rumor spreading in Online social network," in *2012 International Conference on Machine Learning and Cybernetics*, 2012, vol. 4, pp. 1258–1262.

[22] D. J. Daley and D. G. Kendall, "Epidemics and rumours," *Nature*, vol. 204, no. 4963, p. 1118, 1964.

[23] D. J. DALEY and D. G. KENDALL, "Stochastic Rumours," *IMA J. Appl. Math.*, vol. 1, no. 1, pp. 42–55, 1965.

[24] M. Maleki, E. Mead, M. Arani, and N. Agarwal, "Using an Epidemiological Model to Study the Spread of Misinformation during the Black Lives Matter Movement," *arXiv Prepr. arXiv2103.12191*, 2021.

[25] "GitHub - unitaryai/detoxify: Trained models & code to predict toxic comments on all 3 Jigsaw Toxic Comment Challenges. Built using ⚡ Pytorch Lightning and 🤗 Transformers." [Online]. Available: https://github.com/unitaryai/detoxify. [Accessed: 13-Aug-2021].

[26] A. Obadimu, T. Khaund, E. Mead, … T. M.-I. P., and undefined 2021, "Developing a Socio-Computational Approach to Examine Toxicity Propagation and Regulation in COVID-19 Discourse on YouTube," *Elsevier*.