ASEAN Engineering Journal

Full Paper

MODELING THE SPREAD OF FAKE NEWS ON SOCIAL NETWORKING SITES USING THE SYSTEM DYNAMICS APPROACH

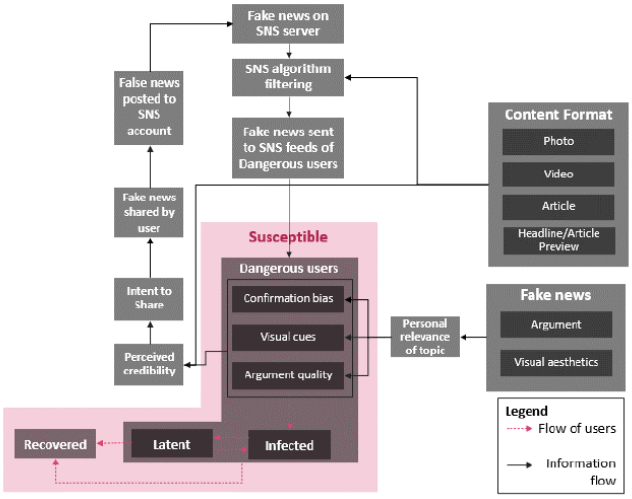
Aleena Concepcion, Charlle Sy*

Industrial & Systems Engineering Department, De La Salle University-Manila, 2401 Taft Avenue, Manila, Philippines, 1004

Article history
Received
12 September 2022
Received in revised form
15 April 2023
Accepted
02 May 2023
Published online
30 November 2023

*Corresponding author
charlle.sy@dlsu.edu.ph

Graphical abstract



Abstract

The problem of false news online has continued to worsen, especially after witnessing significant events around the world unfold, such as the 2018 Cambridge Analytica scandal, COVID-19 pandemic, to the 2021 January 6th Insurrection at the US Capitol. False information online has distorted online users' perception of the real world. As daily life is more intertwined with the digital world, false news becomes a more urgent concern because of the way it can shape public opinion. This study presents a rumor propagation model, which was based on epidemiological models, to address the spread of false news on social networking sites. The existing model was expanded on the STELLA software to consider the cognitive process of users when encountering false news, the platform in which the false news spreads, and the relationship of false news with online users. Simulations showed that Confirmation Bias, Sharing of Posts, and Algorithmic Ranking were the three critical variables of the model. It was found that possible interventions include a mix of reducing the bias of users at a wide-scale level and restructuring the SNS algorithm.

Keywords: System Dynamics, Social Media, Fake News, Disinformation, Simulation

1.0 INTRODUCTION

The diversification of mass media has transformed how news is transmitted and consumed. It is apparent how new media (digital-based content) made news production and distribution faster and more accessible. However, compared to legacy media (print, radio, television), the quality of the news being published through digital means is not controlled. This paved the way for false information to be instantly posted, viewed, and shared by users online. Today, younger generations tend to prefer digital media [1], like social networking sites (SNS), when consuming news. In the early 2000s, SNS like Facebook were launched, allowing users to connect and instantly share content with friends in their network [2]. However, users were only connecting with like-minded users, forming homogenous clusters called echo chambers. These are closed systems that encourage confirmation bias and the validation of pre-existing beliefs [3]. Consequently, false news thrives in environments like these because there is no room for opposing views to challenge the status quo.

Current literature in the propagation of false news is done in line with epidemiological models and rumor propagation models. The most common model used is the classic Susceptible-Infected-Recovered (SIR) where SNS users are the Susceptible, false news sharers as the Infected, and those who fact-check the news as the Recovered [4]. Models have also been further developed to address how the propagation of false information had evolved on the Internet and SNS. An example of this is the Susceptible-Dangerous-Infected-Latent and Recovered (SDILR) rumor propagation model on social networking sites [5]. This

modified SIR model attempted to consider how SNS filters information to users, and how users temporarily forget false news but eventually become "re-infected" when encountering it again. Although this has captured certain dynamics of users being exposed to Infected users in an echo chamber, the model is yet to consider how users directly interact with false news on online platforms, and not just other infected users.

As news consumption shifts to social networking sites, users become more susceptible to false news because of social networks turning into echo chambers [6]. To add to that, digital platforms give way to different content formats like articles, photos, and videos, all of which can be designed to appear as authentic news. With SNS enabling the publishing of unverified information, false news, which is false information designed as authentic news, can be posted, viewed, and shared by users in a matter of minutes. Given how false quickly spread online, it has already been likened to the spread of rumors and epidemics in real life through mathematical models [7].

Existing models on the spread of false news on social networking sites have not fully captured the complex system of false news that involves users' cognitive processes in evaluating information, online social networking sites with filtering algorithms, and their interactions with each other. This results in the formulation of generic propagation models of false information that only focuses on the users of social networking sites.

This study therefore aims to understand how false news spreads on social networking sites 'infects' online users. These will be done by further building on the SDILR model of Yao *et al.,* (2019) [5]. The purpose of this study is to shed light on how users interact with false news on social networking sites. It seeks to capture the cognitive process of online users when exposed to false news, specifically their confirmation bias. The study will also model how false news is behaving on Social Networking Sites and how users subsequently interact with the posts. The rest of the paper is structured as follows: Section 2 provides a literature review on existing modelling approaches on the propagation of fake news. Section 3 presents the system definition that provides the scope of the study. This is followed by the stock flow model development in Section 4. Section 5 contains the computational experiments used to validate the model. This also includes the scenario analysis done to further understand how fake news spread in SNS. Finally, the paper provides a synthesis of the research work and recommendations for further studies in Section 6.

2.0 LITERATURE REVIEW

The SDILR model proposed by Yao *et al.,* (2019) was able to portray the general dynamics of SNS users, how they respond to false news online, and its recurrence [5]. Since SNS have been the main platform where false news is spread, the model must consider how SNS works. In the original SIR model, the Infected state was always preceded by the Susceptible which assumes all SNS users to be equally exposed to Infected users. In the study, a "Dangerous" state was added to represent users who are friends with Infected users. This state creates an echo chamber of users where they become more vulnerable to believing in false news but did not factor in the need for interaction with these users and/or false news itself. The model solely focused on Users' behaviors towards false news online and not necessarily false news posts themselves. In this study, the working definition of Dangerous will be expanded to the state where users are constantly exposed to false news, after interacting with Infected users.

After the Dangerous state, the model allows users to transition to either Infected or Recovered. Users who believe in the false news will be Infected, while those who do otherwise will move to Recovered. The model did not indicate specific drivers of belief for those who get infected.

Following the Infected state is the Latent state. It is described in the study as a "quarantine" period, like a user forgetting a rumor then remembering it when it recurs. In this state, a user can either return to the Infected state if they believe it or become Recovered if they do not. Other studies have considered recurrences of false news by allowing those in the Recovered state to return to the Susceptible state [4]. In this setup, the recurrence is not necessarily captured because it simply treats the susceptible user as someone who is encountering false news for the first time again.

Belief can be driven by several factors like confirmation bias, visual cues, social cues, and medium format. Suntwal *et al.*, (2020) conducted a study to understand what drives people to believe in and share false news [8]. The study found Confirmation bias to be a strong driver of belief. They also found that the veracity of false news is not what drives people to share false news, but confirmation bias and belief. This simply means that people are more likely to share information that aligns with their belief regardless of whether it is true or false.

Apart from Bias, visual cues can also influence users' perceived credibility/belief which in turn is a strong indicator of post engagement [9]. False news heavily relies on the appearance of being real to be considered credible so this study will consider the effect of the design. In Suntwal *et al.,* (2020), the visual cue, Source Likeability, refers to how professional, relatable, and likable the source looks [8]. Similarly, in Fogg et. al (2003) the plurality (46.1%) of over 2,500 respondents ranked the "design look and feel" as the highest indicator for the perceived credibility of news websites [10].

Like website design, the medium in which news is presented online has its effects on the information being delivered. Ireton and Posetti (2018) stated that visuals make people less critical of the news they are consuming [11]. These visuals include photos and videos which are found to significantly influence online users' perceived credibility of a news article [12]. In the study, photos and videos were presented as supporting content to a hypothetical online news article. The content was replaced with placeholder texts and blurred images/videos to isolate the effects of the visual aids. Majority of the studies in the western context focused more on false news articles posted online but in other countries like the Philippines, it is evident that false news is commonly seen in photo and video formats as well. As stated by Szabó (2016), in the digital age, visuals alone are enough to provide information; sometimes this is accompanied by captioned text [13]. On SNS, it has also been observed that images, videos, articles, or video links have varying rates of spreading, and even different audiences for each content type.

SNS are said to have taken the place of mass media as the gatekeepers of information. Specifically, the algorithms that distribute content to newsfeeds. Gillespie (2018) describes its processes of similar to distribution centers where information (posts) are the commodities [14]. Basically, the inputs are user-

generated content, mixed in with advertisements which will be filtered by the algorithm. The output is a personalized news feed with posts ranked in a particular order that maximizes user engagement and revenue. The algorithm predicts and ranks which posts can perform well, and ranks these higher up in news feeds, while there are those that are either ranked low, or not distributed at all. Unfortunately, the factors used by these algorithms are inaccessible to the public. Despite this, general factors that are observable across different SNS can still be considered like affinity, recency, content format, post engagements or popularity [15]. With news feeds being filled with content that are meant to align with a user's pre-existing interests and beliefs, over time, these feeds will become homogenous bubbles. In this environment, confirmation bias will be reinforced, and users are more likely to believe in posts without seeking further validation.

2.1 Propagation Models Of False News

Considering the recent rise of false news on social networking sites, studies on its propagation dynamics have increased as well. The spread of false information has been likened to that of epidemics [16]. The first basic rumor spreading model is the Daley and Kendall (DK) model which was developed in 1964 [17], based on Kermack and McKendrick's Susceptible-Infected-Recovered (SIR) model, both of which are common applications of Ordinary Differential Equations (ODE) [18]. In the DK model, the SIR states were translated to Ignorant-Spreader-Stifler. This was then improved by Maki and Thomson in the MK model by modifying the rate and effect when two Spreaders come into contact, where the two would meet twice the rate of the DK model, and only one Spreader will turn into a Stifler [17]. The SIR model and the two classic rumor propagation models continue to be used in recent studies that focus on social networking sites.

Propagation models adapted from the classic SIR include variations like SI, SIS, SIRS, SEI (Exposed), SEIR, SEIS, SEIRS [19] to study whether recurrences of false news is captured. Fan et al., (2020) modified transmission rates of the SIR model to factor in similarity and popularity of users [7]. Another modified SIR model is The Gossip Model developed by Deters *et al.,* (2019) [18]. The model considered a simple human-to-human transmission of rumors where recovered people can become reinfected after contacting an infected person, and susceptible people can immediately recover if they do not believe in the rumor.

Models derived from the DK/MK propagation have also considered several factors that have been found to contribute to the spread of false news online. The SAIR model [17], like the SI2R model, focused on the bias of the users, but this also considered the behaviors of believing and transmitting, believing and not transmitting, and not believing and not transmitting. An ISCR model from Piqueira *et al.*, (2020) was also made to incorporate those who fact-check the false news so a "Checker" compartment was added [20]. Li *et al.,* (2019) created a homogenous IS2R2 model where a multi-lingual environment was considered, to study the role of language in spreading false news [21].

Since these base models were made before the internet, it was developed in the context of human-to-human transmission. Recent studies have improved on these models to resemble human interaction on online platforms, specifically SNS, by considering echo chambers, algorithm filtering, social cues, and the like. Additionally, because users tend to have different behaviors when browsing online, bias, belief, language, and fact-checking behaviors were factored in.

Although existing models have in many ways considered how users become infected when contacting other users who spread false news, these models have only focused on the users' behaviors. Deters *et al.*, (2019) recommended considering the impact of social networking sites on false news propagation, because direct interaction is no longer necessary in this context [18]. Classical rumor propagation is based on human dynamics before the internet so encountering a spreader is the same as being exposed to false news. However, with online platforms, users need a significant amount of interaction with infected users or false news posts before they can become exposed and infected. Furthermore, existing models have not factored in the cognitive processes of users, like bias, how it is reinforced by false news and influences users' belief. In Hartley and Vu (2020), an equilibrium mathematical model of "fake news" in the context of COVID-19 was developed to model and understand the different factors that influence online users to engage with false news [22]. Although this model acknowledges the presence of SNS, the model failed to capture the relationship of false news posts with the users, by the platform was treated as a simple factor that influences the users' behavior instead of a whole system altogether. Like the other models discussed, the definition of "fake news" in the study was not specified as well. To address these gaps, this study will consider all three: Users, Cognitive Processes, and Online platforms.

3.0 SYSTEM DEFINITION

The system understudy is presented in Figure 1, where each node represents a state that positively influences the following state, and negatively influences the previous state. Other notable relationships in this diagram are the link from Dangerous to Recovered, and the returning link of Latent to Infected (in blue). The former represents users who become exposed to false news but do not get Infected, and the latter are those users who become re-Infected when the false news recurs online.

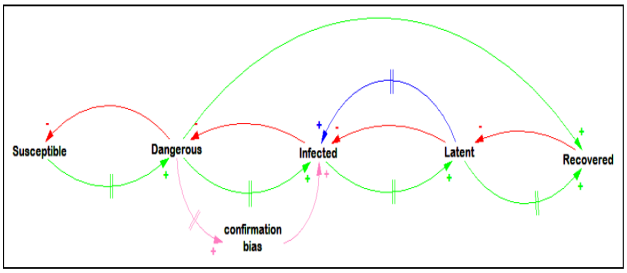


Figure 1 False news propagation with delay and confirmation bias

The system also considers delays in different links, and the role of confirmation bias. In the original SDILR model, a fraction of Susceptible users can immediately become Dangerous after encountering Infected users. On SNS platforms, a single encounter is not deemed necessary. The algorithm needs a significant amount of interaction between Infected and Susceptible before it can determine that the Susceptible user is

interested in false news. In the diagram, a delay mark denoted by double strikethrough lines is placed to represent this. Likewise, users from Dangerous to Infected also need time to be continuously exposed to false news so that their bias will also be reinforced and compound. In the next two links from Infected to Latent, and vice versa, delays are placed to account for forgetting and recalling as both are related to the passing of time and the degree of the issue in the media (ex. how often is it talked about or is it still a trending topic). Lastly, the link from Latent to Recovery also requires a delay because existing corrective measures like fact-checking are done at a certain point in time after the false news was posted.

4.0 MODEL DEVELOPMENT

The System Dynamics (SD) modelling methodology utilizes the stock flow diagram to model and provide quantification for the relationships that exist within the system. The diagram is translated into integral equations, usually facilitated through the use of high-level simulation programs, such as STELLA that is used in this study. System variables are generally classified according to two types: stocks and flows. Stock variables represent the accumulations in the system. As mentioned by Sterman (2000), they characterize the state of the system and generate the information upon which decisions and actions are based [23]. Stocks create delays by accumulating the difference between the inflow to a process and its outflow. These variables determine the state of the system and are dependent on past values. In contrast, flows are unable to accumulate through time. They simply alter the quantity of the stocks by being either an inflow or an outflow to it [24]. This quantity is determined by summating the inflows less the outflows of a particular stock.

The system understudy will be divided into two sub-models: Users and Online Platform. The first sub-model (Users) is the base SDILR model adopted from Yao *et al.,* (2019) [5]. This will represent the people or users being "Infected". The base model has been extended to also include the cognitive process of users when seeing false news. The second sub-model, Online Platform, will focus on the system where the "virus" or false news is distributed, regulated, and encountered.

4.1 User Sub-Model

The sub-model, shown in Figure 2, starts with a constant inflow of users to the Susceptible stock, with a birth rate equal to the death rate. From there, users transition to Dangerous when they come into contact with Infected users. The Dangerous state of the base model was defined to be where users are friends with Infected users or those who share false news. There are two outflows in the Dangerous stock, one for when users do not believe in the false news and become Recovered, and another for when they do believe and become Infected.

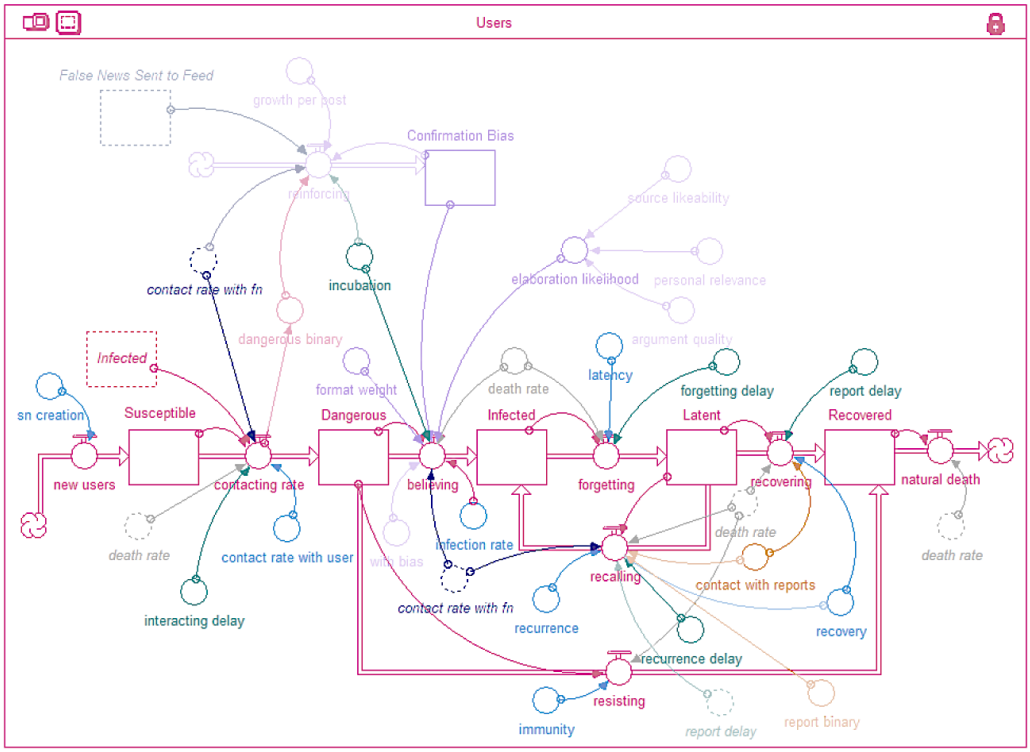


Figure 2 Stock flow diagram for user sub-model

When users do become Infected, they eventually forget the false news and become Latent. In the Latent stock, users can either Recover or return to Infected when the false news recurs. Each stock also has a natural death rate outflow. The general equation for each flow is simply the connected stock/s multiplied by the corresponding rate of each flow, minus the natural death rate multiplied by the stock. Infected users have an additional inflow of users returning from Latent multiplied by the recurrence rate.

Four types of delays (highlighted in green in Figure 2) were considered to capture different events that take place when

users transition from one state to another. These are defined as interacting, incubation, recurrence, and reporting delays. An interacting delay is needed when Infected users come into contact with Susceptible users. This allows users to interact with Infected users to a point where the algorithm deems their affinity as significant to the Susceptible users. After this delay, users transition to Dangerous.

An Incubation delay is placed in the flow from Dangerous to Infected. This study defines the Dangerous stock as the state wherein users are being exposed to false news. The delay allows users to first be constantly exposed to false news for their biases to be reinforced, which would lead them to believe in the false news and become Infected. A forgetting delay is placed from Infected to Latent since it takes time for a current event to become latent and be forgotten by a user. Likewise, the recurrence of false news occurs after a period of latency, hence the recurrence delay in the returning flow from Latent to Infected. Lastly, in the flow from Latent to Recovery, a reporting delay is added to account for the time it takes for the media or other information sources to respond and release a statement that debunks the false news.

A loop for confirmation bias is added to consider how continuous exposure to false news reinforces one's bias. In this loop, the amount of false news sent to the news feeds is multiplied by the amount of growth in bias each post contributes, as well as the contact rate of a Dangerous user with false news. A binary variable connected to the contacting rate is included here to signal the bias loop to be switched on once Susceptible users are contacting Infected users. The same incubation delay for believing is used in reinforcing since these occur simultaneously, while users are being exposed to false news. In the interest of brevity, the discussion concerning the auxiliary variables and the actual equations used in the model can be found in the Appendix.

4.2 Online Platform Sub-Model

The second sector of the model involves the sub-system of social networking sites where the users come into contact and eventually interact with false news. The stock flow diagram for the sub-model is shown in Figure 3.

The distribution of posts on SNS starts with people creating false news posts (*fn creation*) that are sent to the servers of SNS. Aside from the constant creation rate of false news, the act of sharing false news also contribute to the number of posts online. These "shares" are converted to the number of posts by multiplying the sharing rate to the average depth of a share (how many more times a share is reposted). The posts in the stock False News Online can be sent to news feeds, deleted, or stay in the stock.

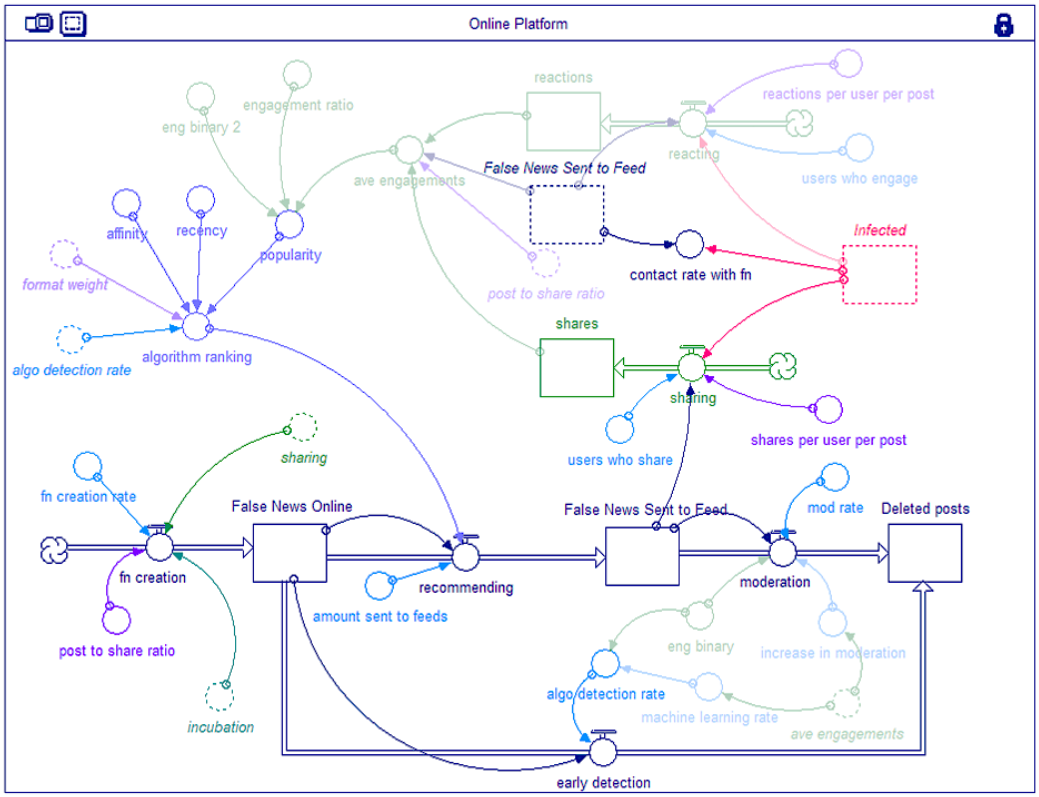


Figure 3 Stock flow diagram for online platform sub-model

Today, most SNS have their algorithmic detection system that can downrank or delete posts if they go against community standards. The platform's recommendation system will begin to rank and filter posts to send on News Feeds. Since there is a large amount of content posted every day, the recommendation system can only recommend a small number of posts to news feeds and prioritizes those with higher rankings. This explains how posts can remain in the stock of False News Online as posts that were not recommended and not deleted. There is a short delay in creating false news, while the sharing of posts only

occurs after the incubation period of dangerous users. Recommending and early detection do not have delays since these are automated processes. Moderation will also have delays to allow a waiting time for post-engagement to accumulate and a reviewing time for human reviewers.

The diagram also contains the dynamics involved in algorithm ranking. This variable consists of the format weight, affinity, recency, and popularity of posts. Posts with higher rankings appear higher in News Feeds, increasing the chances of the post being seen by users. In the base run, the algorithm ranking variables are set as constants. Meanwhile, the Sharing flow allows Infected users share false news posts. The shares are computed by multiplying the Infected users and the False News Sent to Feed by a ratio of shares per user per post. Since not all Infected users share false news posts, there is also a constant variable to represent the percentage of infected users who share.

5.0 RESULTS AND DISCUSSION

The reference mode to be used in this study will be based on the dataset used by Vosoughi *et al*. (2018) [25]. The dataset contains English language tweets collected from 2008 to 2016 with varying degrees of veracity (True, False, Mixed) on different topics (Politics, Business, Science and Tech., etc.) and claims. Each tweet was labeled with a topic and the claim it pertains to so a case study on a single claim with several posts regarding the claim can be done. However, the specific claim being discussed in the set of tweets is not further specified. It is necessary to focus on a single claim so that the behaviors of other claims do not clash with the collected data (behavior of claims about the elections may differ from claims about day-to-day political news). Furthermore, distinct users and tweets were identified with a User and Tweet ID respectively, along with a timestamp. Retweeted posts and other interactions on a post were also labeled accordingly.

The graph of the Infected users was derived from computing for the distinct number of users, retweets, and interactions. On Twitter, users can only retweet or react to a post once so there is a 1:1 ratio of interaction to user ratio, whereas, on Facebook, users can share a post multiple times but react on the original post once. Likewise, the graph for False News Sent to News Feeds is simply the sum of unique Tweet IDs and retweets. The resulting graphs, presented in Figure 4, are indicative of how false news and the number of infected users build up and recur over long periods. This dataset however is only limited to recorded timestamps on tweets, which explains the simultaneous behavior of the two variables.

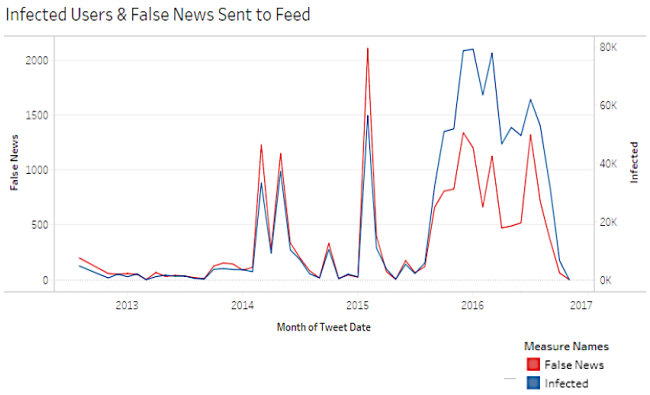


Figure 4 Reference modes for infected users and number of false news

5.1 Base Run

In this section, the final stock flow model will be simulated within a 10,000-time unit horizon using the STELLA software. Figure 5 shows how Infected and False News Sent to Feed interact with each other under the simulation run. When the online platform was connected to the online users the behavior of infected users is similar to that of the reference mode, as well as ones described in the literature, even with the algorithm ranking set with constant values. It is easily seen how false news and users are reactive to each other since the peaks of each cycle do not intersect at the same time. It shows that there is a delay for the false news to be picked up by users and be recommended to them. Likewise, there is a delay for false news to go down after users become latent because the algorithm needs time to react to the interaction of the users.

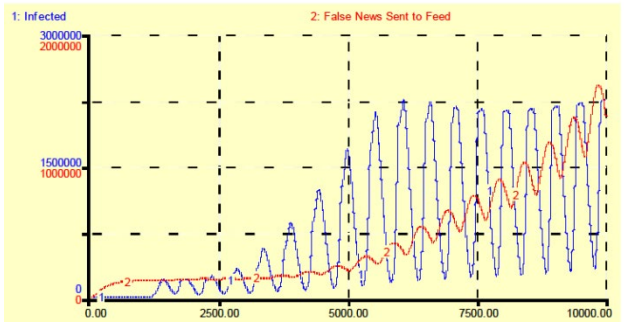


Figure 5 Infected and False new sent to feed

Interestingly, the graphs on Figure 6 show how false news can drive the recurrence of infection because as Infected users transition to Latency, the online platforms are still recommending false news to feeds. Just when Infected users have already forgotten the false news, the platforms increase recommending and triggers the recalling of the already Latent users. This reflects the literature on the structures of algorithms because they are structured to drive users to engage with posts thus recommended users with content that they have previously liked, in an attempt to gain user engagement.

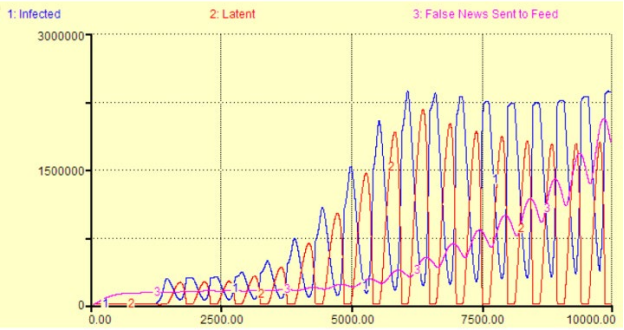


Figure 6 Infected, Latent, and False news sent to feed

In the early stages before t=5000, Figure 7 and Figure 8 show that when there is a relatively constant inflow of false news, only a small number of users become infected because most of the users are still in the dangerous state and confirmation bias is still low ranging from the initial value to 0.07 This is also consistent with studies saying that false news requires constant repetition to gradually reinforce bias and be planted in the minds of users. At around t=6000, the cycles of the oscillation start to remain within the same range. By the end of the simulation, the bias level becomes extremely high but still yielded the same number of infected users at t=6000 when bias was relatively lower. This finding suggests that the timing of when interventions occur is critical because if bias is already high by then, it might be difficult to change the minds of the users.

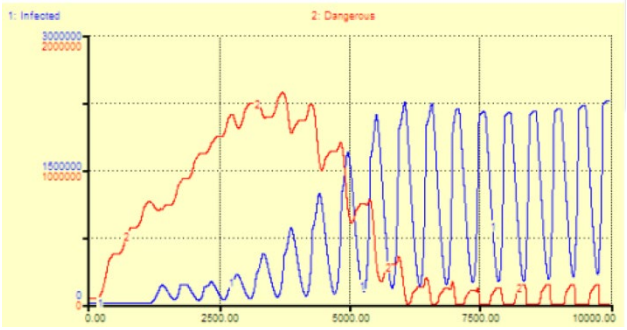


Figure 7 Infected vs Dangerous

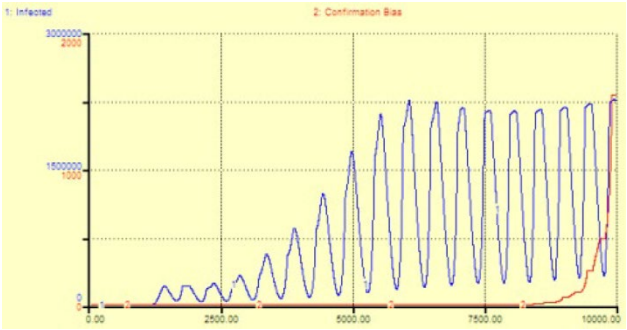


Figure 8 Infected vs Confirmation bias

5.2 Sensitivity Analysis

A sensitivity analysis was performed to further observe the behavior of the model and its reaction to input parameter changes. Subsequent simulation runs modified model parameters to increase or decrease by 50%. The resulting graphs consist of three runs: the base run of the model colored in blue (Run 1), the run with a 50% decrease in red (Run 2), and the run with a 50% increase in pink (Run 3).

The first parameter tested was the recurrence rate. When the base model was analyzed, it was found that the model was not sensitive to the recurrence rate. However, after having expanded the model, the model became sensitive to the change in the parameter (see Figure 9). When the recurrence rate is lower, the oscillations slowly decrease in amplitude because fewer latent users are recalling the news. However, when it is higher, the amplitude continues to increase over time since the claim is resurfacing on the users' fields.

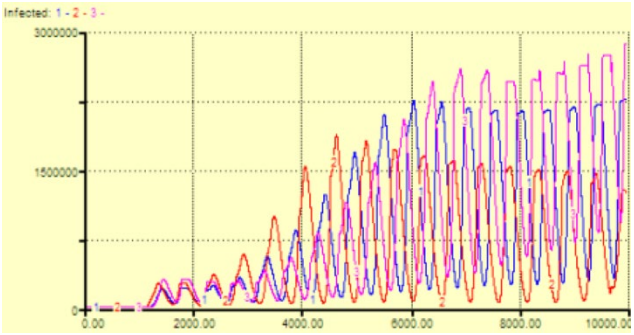


Figure 9 Effects on the infected from changes in the recurrence rate

Figure 10 shows that when recurrence is lower, the amount of false news sent to feeds is less than the base run. When it is higher, it increases at a faster rate. These findings suggest that the recurrence of false news after the claim has died down should be limited so that the latent users can transition to recovery instead.

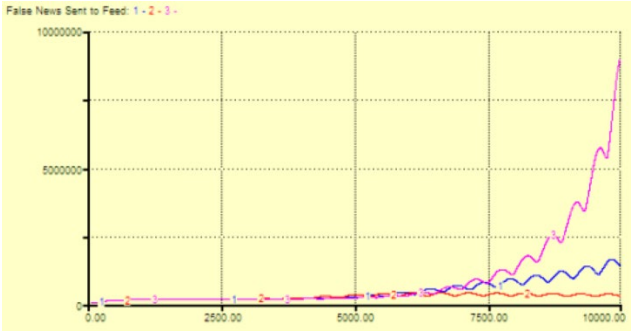


Figure 10 Effects on the infected from changes in the recurrence rate

Apart from these findings, this behavior change also aligns with the media theory of Agenda Setting where media institutions have the power to create salience by the mere selection of topics and frequency of covering such topics [26]. In other words, when a certain topic or claim is continuously discussed in the media, a sense of importance and/or urgency is created. This concept is often associated with mass media to exhibit how organizations can shift public attention and determine what the public should focus on. However, when this is applied to Social Networking Sites, the salience of a topic only exists within the news feed of each person because of the filtering algorithm. When a user actively engages with false news, these posts are recommended over and over, the users will view the false news as something urgent and eventually become reinfected.

The effect of changing the degree of confirmation bias was also tested. The resulting graphs for Infected users and False News Sent to Feed are shown in Figure 11 and Figure 12, respectively.

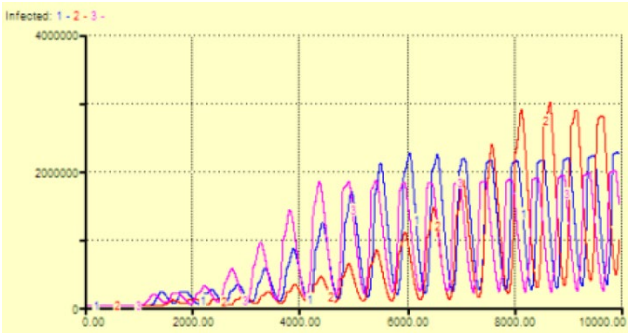


Figure 11 Effects on the infected from changes in the confirmation bias

Interestingly, decreasing the growth of bias per post only delayed the increase in both infected users and false news. When the variable was increased, infection shifted earlier, with lower peaks than the base run and the second run, while the amount of false news was unexpectedly lower than the two runs. In this case, it is likely that the infected users were distributed earlier when there was still a small number of false news since the bias threshold was reached earlier as well. This simply proves that when bias is initially high, less false news is needed to Infect the user. These results also acknowledge how people are inherently biased— even if the growth rate is smaller, when it is continuously reinforced there will eventually come a time when it exponentially grows. A solution to address reinforced confirmation bias is not to reduce the growth rate, but to counter the accumulated bias earlier instead.

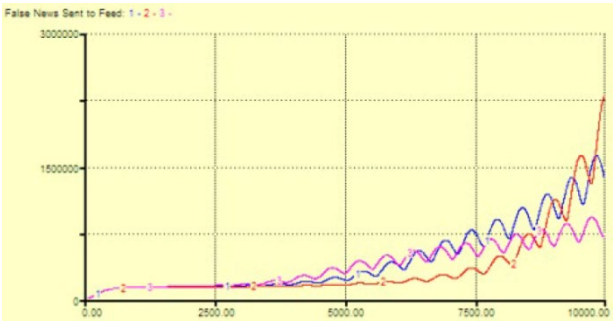


Figure 12 False news sent to feed under changes in confirmation bias

5.3 Scenario Analysis

A scenario was simulated where the algorithm recommendation is based on post popularity to study the behavior of users and false news with at least one variable algorithm factor (see Figure 13). A reacting flow where a fraction of infected users reacts to the false news, similar to the sharing flow, was added to the model. This makes the algorithm prioritize posts that are receiving engagements from users, with shares weighed heavier than reactions.

Figure 14 shows the resulting graph of Infected users after considering popularity. Compared to the base run, it took longer for Infection to occur since the post still needs to gain popularity rankings. Despite the slow start, the Infection began to grow exponentially while the amplitude of each cycle decreases and plateaus at the end. Meanwhile, the behavior of false news in Figure 15 no longer oscillated, but simply grew exponentially. This suggests that over time, users get stuck in the Infected state because the algorithm will keep on recommending the false news that is guaranteed to elicit activity from the users. Another scenario with user affinity was also simulated, which also resulted to the same behaviors. This proves that the current architecture of SNS algorithm, where user engagement is optimized, can easily be used and exploited to circulate false news.

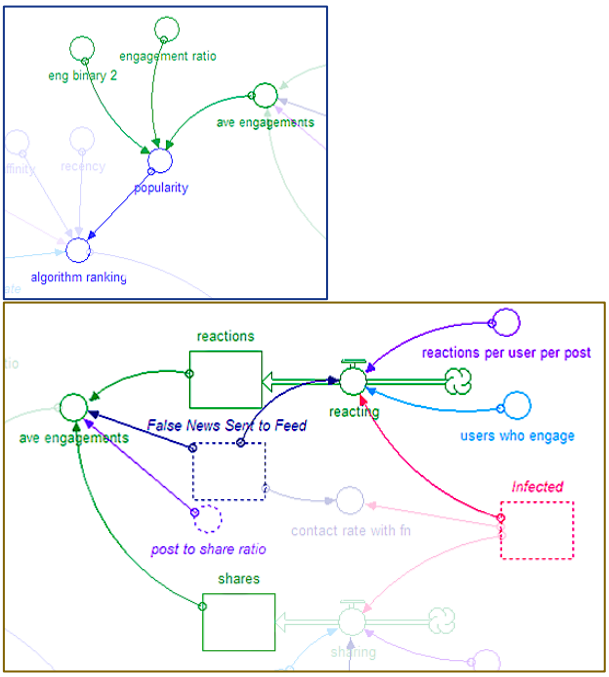


Figure 13 Popularity based on post engagement loop

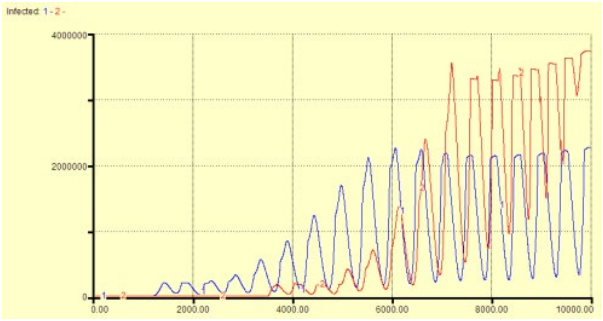


Figure 14 Effects of popularity on number of infected

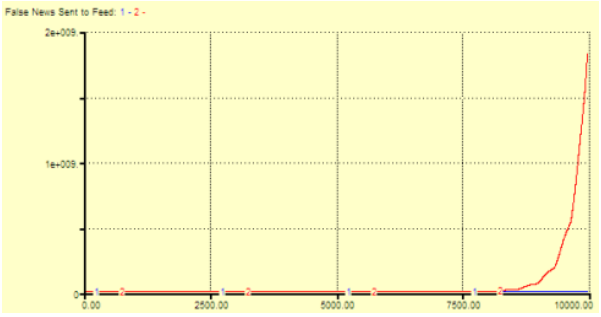


Figure 15 Effects of popularity on false news

6.0 CONCLUSION

As daily life became more mediatized over the years, false news online has only continuously increased. This study aimed to model this spread on social networking sites by adopting a rumor propagation model. The model was extended to include the dynamics involved in the online platforms themselves and the users' cognitive processes. Specifically, the model depicted how the behavior of online users have inadvertently created echo chambers that had in turn enabled false news to flourish online. This has been confirmed in the base run analysis when it was found that confirmation bias and the actual sharing of online content represent critical variables of the system. This signifies that interventions should focus on targeting the dynamics in the loops to which these variables are present.

Future research could focus on developing policies that could target the spread of false news. The model could also be further refined in terms of its accuracy and include additional user dynamics. The study was not able to consider the specific factors that are used in SNS algorithms due to its inaccessibility, especially the influence of social cues, user interface, bias towards the news sources, etc. Moreover, this study was only limited to users on one SNS platform but today, it is more common that Users have more than one SNS account which could have different dynamics across different platforms too. It is also recommended to consider the behavior of false news in different cultural contexts, like false news in bilingual or multi-lingual countries. Future studies could also transition into developing strategies and policies that would minimize the spread of false news and information online.

Acknowledgement

The authors are grateful for the funding provided by De La Salle University – Manila, which allowed for the completion of this study.

References

[1] Schroeder, R. 2018. Media systems, digital media and politics, *Social Theory after the Internet*. 28–59.

[2] McNair, B. 2017. The decline of trust in journalism, In *Fake News: Falsehood, Fabrication and Fantasy in Journalism*. Routledge, London.

[3] Törnberg, P. 2018. Echo chambers and viral misinformation: Modeling fake news as complex contagion, *PLoS ONE*. 13(9): 1–21.

[4] Oktaviansyah, E. and Rahman, A. 2020. Predicting hoax spread in Indonesia using SIRS model, *Journal of Physics: Conference Series*. 1490(1): 1-5. DOI: https://doi.org/10.1088/1742-6596/1490/1/012059.

[5] Yao, Y., Xiao, X., Zhang, C., Dou, C., and Xia, S. 2019. Stability analysis of an SDILR model based on rumor recurrence on social media, *Physica A: Statistical Mechanics and Its Applications*. 535: 122236. DOI: https://doi.org/10.1016/j.physa.2019.122236.

[6] Campan, A., Cuzzocrea, A. and Truta, T. 2017. Fighting fake news spread in online social networks: actual trends and future research directions. *IEEE International Conference on Big Data*. 4453–4457. DOI: https://doi.org/9781538627150.

[7] Fan, D., Jiang, G. P., Song, Y. R. and Li, Y. W. 2020. Novel fake news spreading model with similarity on PSO-based networks. *Physica A: Statistical Mechanics and Its Applications*. 549: 124319. DOI: https://doi.org/10.1016/j.physa.2020.124319.

[8] Suntwal, S., Brown, S., and Patton, M. 2020. How does Information Spread? An Exploratory Study of True and Fake News, *Proceedings of the 53rd Hawaii International Conference on System Sciences*. 3: 5893–5902.

[9] Oyibo, K., Adaji, I., Orji, R., and Vassileva, J. 2018. What drives the perceived credibility of mobile websites: classical or expressive aesthetics? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 10902 LNCS(1): 576–594.

[10] Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. 2003. How do users evaluate the credibility of Websites?: A study with over 2,500 participants, *Proceedings of the 2003 Conference on Designing for User Experiences, DUX '03*. 1–15.

[11] Ireton, C. and Posetti, J. 2018. *Journalism, 'Fake News' & Disinformation,* UNESCO, Paris.

[12] Wobbrock, J. O., Hsu, A. K., Burger, M. A., and Magee, M. J. 2019. Isolating the effects of web page visual appearance on the perceived credibility of online news among college students, *HT 2019 - Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 191–200, DOI: https://doi.org/10.1145/3342220.3343663.

[13] Szabó, K. 2016. Online Visuality. In A. Benedek & Á. Veszelszki (Eds.), *In The Beginning was the Image: The Omnipresence of Pictures,* 103-150, Peter Lang, New York.

[14] Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven.

[15] Lara-Navarra, P., López-Borrull, A., Sánchez-Navarro, J., and Yànez, P. 2018. Medición de la influencia de usuarios en redes sociales: *Propuesta socialengagement. Profesional de La Informacion*. 27(4): 899–908.

[16] Meel, P. and Vishwakarma, D. K. 2020. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*. 153: 112986.

[17] Zhu, L. and Wang, B. 2020. Stability analysis of a SAIR rumor spreading model with control strategies in online social networks. *Information Sciences*. 526: 1–19.

[18] Deters, J., Aguiar, I., and Feuerborn, J. 2019. The Mathematics of Gossip. *CODEE Journal*. 12(1): 73–82.

[19] Piqueira, J. R. C., Zilbovicius, M., and Batistela, C. M. 2020. Daley–Kendal models in fake-news scenario. *Physica A: Statistical Mechanics and Its Applications*. 548: 123406.

[20] Li, J., Jiang, H., Yu, Z., and Hu, C. 2019. Dynamical analysis of rumor spreading model in homogeneous complex networks. *Applied Mathematics and Computation*. 359: 374–385.

[21] Hartley, K. and Vu, M. K. 2020. Fighting fake news in the COVID-19 era: policy insights from an equilibrium model. *Policy Sciences*. 53(4): 735–758.

[22] Sterman, J. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*, Irwin McGraw-Hill, Boston.

[23] Sy, C. 2017. A policy development model for reducing bullwhips in hybrid production-distribution systems, *International Journal of Production Economics*. 190: 67-79.

[24] Vosoughi, S., Roy, D., and Aral, S. 2018. The Spread of True and False News Online, *Science*. 1151: 1146–1151.

[25] McCombs, M. 2011. *The Agenda-Setting Role of the Mass Media in the Shaping of Public Opinion*, University of Texas: Austin, https://doi.org/10.13245/j.hust.15S1016.

Appendix

Appendix A. Auxiliary variables in the user sub-model

Contact rate with false news - In the flows contacting rate, believing, and recalling, the variable contact rate with false news will be multiplied to consider the probability of users encountering false news. For the contacting rate flow, only a fraction of the contact rate with false news will be considered since they are not yet considered as users who are interested in false news.

Confirmation Bias - A stock for confirmation bias is connected to the believing flow, which will be multiplied by the users who are transitioning from Dangerous to Infected. In the equation for believing, Confirmation bias will not be included within the incubation delay because the effect of bias is instinctive and immediate.

Elaboration likelihood - A variable for elaboration likelihood is also connected to believing to represent the cognitive processes that users take when presented with information. This variable determines which route users take based on the personal relevance of the topic to them. Based on the literature, when personal relevance is low the user will more likely elaborate based on heuristic cues like source likeability. When personal relevance is high, the user will more likely depend on the argument quality. This variable will simply be multiplied by the believing rate. Like confirmation bias, this factor will come after the delay.

Format weight - As discussed earlier, the types of formats affect the believability of false news. Generally, more visual types of formats like photos and videos are more believable than text-based formats like articles, tweets, and headlines. The more visual a post is, the more it is believed. This variable will also come after the delay since information processing happens instantly on SNS.

Contact with reports - The recovering flow will consider the probability in which users encounter corrective reports because not all infected or latent users are guaranteed to see these reports, especially with the recommendation system of SNS.

Dangerous(t) = Dangerous(t - dt) + (contacting - believing - resisting) * dt
INIT Dangerous = 30000
INFLOWS:
contacting = Susceptible*Infected*contact_rate-Susceptible*death_rate
OUTFLOWS:
believing = Dangerous*infection_rate-Dangerous*death_rate
resisting = Dangerous*immunity-death_rate*Dangerous
Infected(t) = Infected(t - dt) + (believing + recalling - forgetting) * dt
INIT Infected = 30000
INFLOWS:
believing = Dangerous*infection_rate-Dangerous*death_rate
recalling = Latent*recurrence-death_rate*Latent
OUTFLOWS:
forgetting = Infected*latency-death_rate*Infected
Latent(t) = Latent(t - dt) + (forgetting - recalling - recovering) * dt
INIT Latent = 0
INFLOWS:
forgetting = Infected*latency-death_rate*Infected
OUTFLOWS:
recalling = Latent*recurrence-death_rate*Latent
recovering = Latent*recovery-Latent*death_rate
Recovered(t) = Recovered(t - dt) + (resisting + recovering - natural_death) * dt
INIT Recovered = 0
INFLOWS:
resisting = Dangerous*immunity-death_rate*Dangerous
recovering = Latent*recovery-Latent*death_rate
OUTFLOWS:
natural_death = death_rate*Recovered
Susceptible(t) = Susceptible(t - dt) + (births - contacting) * dt
INIT Susceptible = 40000
INFLOWS:
births = Population*death_rate
OUTFLOWS:
contacting = Susceptible*Infected*contact_rate-Susceptible*death_rate

Figure 16 STELLA equations for user sub-model

Appendix B. Auxiliary variables in the online platform sub-model

Algorithm Factors - The algorithm ranking variable consists of the format weight, affinity, recency, and popularity of posts. Posts with higher rankings appear higher in News Feeds, increasing the chances of the post being seen by users. The equation for the rankings will simply be the products of the format weight, affinity, recency, and popularity, and the complement of the algorithmic detection rate.

Sharing of False News - The final stock and flow to be added is the sharing part of SNS. This flow is responsible for making infected users share false news posts. The shares are computed by multiplying the Infected users and the False News Sent to Feed by a ratio of shares per user per post. Since not all Infected users share false news posts, there is also a constant variable to represent the percentage of infected users who share.

Online Platform
Deleted_posts(t) = Deleted_posts(t - dt) + (moderation + early_detection) * dt
INIT Deleted_posts = 0
INFLOWS:
moderation = DELAY(False_News_Sent_to_Feed*mod_rate,150,0)*(1-eng_binary)+ DELAY(False_News_Sent_to_Feed*mod_rate*increase_in_moderation,150,0)*eng_binary
early_detection = algo_detection_rate*False_News_Online
False_News_Online(t) = False_News_Online(t - dt) + (fn_creation - recommending - early_detection) * dt
INIT False_News_Online = 315
INFLOWS:
fn_creation = (DELAY(fn_creation_rate,50)*(1-trust_in_news)+DELAY(sharing*post_to_share_ratio,incubation))
OUTFLOWS:
recommending = False_News_Online*amount_sent_to_feeds*algorithm_ranking
early_detection = algo_detection_rate*False_News_Online
False_News_Sent_to_Feed(t) = False_News_Sent_to_Feed(t - dt) + (recommending - moderation) * dt
INIT False_News_Sent_to_Feed = 173
INFLOWS:
recommending = False_News_Online*amount_sent_to_feeds*algorithm_ranking
OUTFLOWS:
moderation = DELAY(False_News_Sent_to_Feed*mod_rate,150,0)*(1-eng_binary)+ DELAY(False_News_Sent_to_Feed*mod_rate*increase_in_moderation,150,0)*eng_binary
reactions(t) = reactions(t - dt) + (reacting) * dt
INIT reactions = 4548
INFLOWS:
reacting = False_News_Sent_to_Feed*Infected*reactions_per_user_per_post*users_who_engage
shares(t) = shares(t - dt) + (sharing) * dt
INIT shares = 20
INFLOWS:
sharing = shares_per_user_per_post*Infected*users_who_share*False_News_Sent_to_Feed

Figure 17 STELLA equations for online platform