TwitterSim: A policy-oriented test-bed for the spread of Contentious Messages in Twitter

Ramon Villa-Cox[1] and Kathleen M. Carley[1]

Carnegie Mellon University, Pittsburgh PA 15213, USA

1 Introduction

There has been a renewed interest in the spread of disinformation on online social media and its effects on public discourse. In academia, much of the effort has been spent characterizing the salient traits of the diffusion process of false information, especially as it relates to other types of information. However, much work remains to be done to explore viable avenues for policy that would curtail the spread of disinformation or mitigate its pervasive impacts. [3], posit two venues of intervention, those that seek to educate and empower individuals in their interactions with social media and those that promote structural changes in these platforms that limit users exposure to false information. Some work has been done to explore the former, by investigating community-based options for fact checking and reporting bad actors [6]. However, to the best of our knowledge, no work has yet to explore effective ways to design policies that implement these interventions with consideration of the target social network ecosystem.

The first stage of our work is concerned with the collection of relevant instances of community calling-out of contentious or rumorous statements to serve as validation case studies. For this purpose, we are developing machine leaning algorithms for the automatic detection of the stance (support, deny or comment) of Twitter responses to identify contentious or possibly rumorous source tweets in different events. The second stage is concerned with the construction of a viable test-bed for policies that seek to increase community resilience to these types of messages. We propose to design an agent-based dynamic-network model [2] that, by focusing on Twitter, will explicitly model the ways of interaction allowed in the platform and the way that content is broadcasted between users. In this way, we can evaluate the effect of the platform's architecture on the diffusion process of different types of stories. Possible policy implications can center on changing the way that content is shown to give access to opposing views or that amplify messages debunking false stories.

2 Background and Related Work

Relation-based argument mining The first stage of this research is concerned with the identification the relations between argumentative texts. This can be a hard task as it may require knowledge of both the argumentative content and logical structure of different claims [4]. However, a growing number

of researchers have attempted to use argument mining to automatically detect rumors, disinformation, or polarized discussions on social networking sites. For example, [8] exploit the sequential nature of conversation threads on Twitter to augment the relational classification task. Although growing work in this area suggests its significance as a computational problem, in the context of social media, the state-of-the-art predictive performance remains less than ideal [4].

Diffusion of contentious messages There have been numerous studies attempting to characterize the diffusion process of messages, in both speed and scale, on online social media. Due to its public nature, most of these studies have concentrated on analyzing behavior in Twitter. From this literature, we distinguish two main bodies of work, one concerning a system level (top-down) approach and those that take a bottom-up perspective. Our work concerns the later, which relies on Agent-Based Models (ABMs), a popular approach for the modelling of complex adaptive social and natural systems. In the context of information diffusion and rumor spreading in online social media, models have attempted to capture the fundamental mechanisms behind the spread of these messages. For example, [5] implemented an information diffusion model [6] as an ABM, testing whether the anti-rumor spreading behavior of recovered users was reflected in empirical data. They find that these users will typically not affect the recovery of their neighbors, challenging the view that community-based self-correction can occur without further intervention.

3 Methods

Twitter Simulation Although the models discussed have attempted to model information diffusion in an online medium (mostly Twitter), they generally do not consider the effect that the platform has on the formation of user networks and their interaction. These environments, however, impose a series of rules for interaction and content promotion that have an important effect on the diffusion of information. For this reason, it is important that simulations incorporate the structural properties of the platform studied as well as mechanisms for individual user behavior. This allows the evaluation and design of policies that seek to impose structural changes to the platforms in order to hamper the spread of disinformation or that increase user resilience. The proposed simulation will incorporate this logic by carefully implementing the ways that users can communicate in Twitter and the broadcasting rules for these messages. Similarly to [7], each users timeline will modeled as a queue, where new tweets are placed at the top and older ones are removed. As mentioned in their paper, this introduces an economy of attention, which imposes competition among different topics discussed in the environment. We consider the three possible forms of response available in Twitter: retweets, direct replies and quotes and the different ways that these messages are broadcasted to other users. In order to make use of these different types of interaction, it is not sufficient for the cognitive model of users to include knowledge of a subset of the discussed topics, but it also

needs to include a stance to each of them. In this way, the incentives of a user to respond positively (negatively), increase as her beliefs correlate more positively (negatively) with the relevant message. In this we will deviate from the standard model of the timeline (as implemented in [7]), where all messages have a uniform probability of being replied to. Users also will have the capacity to modify their social network by following/unfollowing other users. These decisions will be modelled based on the mechanisms established in the Construct model for learning [2], but adapting them the context of online social media.

Validation Strategy Different model parameters, like frequency of interaction and user propensity to tweet and respond based on stance towards a message, will be instantiated based on empirical estimation from different case studies with known rumorous stories. For example, in our previous work [1], three different types of false stories were identified on Twitter discussions. Here, the diffusion process of these stories, reactions to them and underlying communities were explored. Moreover, we are currently working on a neural network classifier to label the stance of Twitter responses in order to automatically identify contentious or rumorous tweets. We seek to improve on the state of the art in two ways, by including an attention mechanism over conversation threads, to increase accuracy over longer threads, and by designing a collection methodology oriented towards denials (which is considerably under-sampled in available datasets [4]). This tool will serve to expand the available case studies for validation of the proposed simulation. For the validation, we will evaluate the simulated diffusion process of the different types of stories against what is observed on the different case studies. Similarly, the heterogeneity of discussed topics, predicted network topologies, diffusion of positive and negative responses can be contrasted against empirical data.

References

1. Babcock, M., Villa-Cox, R., Kumar, S.: Diffusion of pro-and anti-false information tweets: the black panther movie case. CMOT **25**(1) (2019)
2. Carley, K.M., Martin, M.K., Hirshman, B.R.: The etiology of social change. Topics in Cognitive Science **1**(4), 621–650 (2009)
3. Lazer, D.M., Baum, M.A., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
4. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proc. of SemEval. pp. 31–41 (2016)
5. Serrano, E., Iglesias, C.Á., Garijo, M.: A novel agent-based rumor spreading model in twitter. In: Proc. of WWW. pp. 811–814. ACM (2015)
6. Tripathy, R.M., Bagchi, A., Mehta, S.: A study of rumor control strategies on social networks. In: Proc. of ACM. pp. 1817–1820. ACM (2010)
7. Weng, L., Flammini, A., Vespignani, A., Menczer, F.: Competition among memes in a world with limited attention. Scientific reports **2**, 335 (2012)
8. Zubiaga, A., et al.: Discourse-aware rumour stance classification in social media using sequential classifiers. Information Processing & Management **54**(2) (2018)