The Firestarting Troll, and Designing for Abusability

Andrew Beers,[1] Sarah Nguyễn,[2] Maya Sioson,[1] Mariam Mayanja,[3]
Monica Ionescu,[1] Emma S. Spiro,[2] Kate Starbird,[1]
University of Washington, [1]Department of Human Centered Design and Engineering, [2]Information School,
[3]Paul G. Allen School of Computer Science and Engineering
albeers@uw.edu, snguye@uw.edu

Abstract

The COVID-19 pandemic has provided ample opportunity for the spread of misinformation and bigotry in online debates. In this paper, we use qualitative and quantitative methods to profile a single user whose unique posting methods and messaging contributed to an outsize impact on COVID-19 discourse. We use a dataset of reply threads in response to United States governors, who often announced updates and regulations relating to the COVID-19 pandemic. The user we identified had a highly unusual ability to generate high levels replies from other users disputing and supporting their posts, while rarely engaging in argument themselves and having few followers. We term this user's behaviors as firestarting, which we define to be the goal of starting bad faith arguments without any subsequent participation. To address such behaviors, we suggest a framework of designing for abusability, which focuses on reducing usability for some users in order to improve usability for most users.

Introduction

From March 2020 to October 2020, one Twitter account launched a relentless trolling campaign against 50 United States (US) governors. This account, whose username was a common English name and who we will pseudonymously refer to as James, posted 16 hours a day, seven days a week, as many as 300 times a day, and never fewer than 10. As US governors announced various regulations meant to stop the spread of COVID-19, James systematically replied to each with misleading and antagonistic claims about the dangers of the disease, sometimes posting the same comment to each governor spread out over an 18 hour period. Their comments sparked an enormous amount of argument in subsequent replies, antagonizing the governors' supporters and rallying the governors' opponents. James had less than 1,000 followers, no profile information, a stock image for a profile photo, and rarely engaged in substantive follow-up discussion. Nevertheless, James' posts were so numerous and combustible, that more than 1 out of every 200 tweets arguing about the governors' posts during this time were contesting one of James' many spurious claims.

Many researchers have studied the effects on discourse of the social media influencer, broadly defined as a popular user with outsize influence on social media discourse (Bakshy et al. 2011). Especially in the last year, studies have drawn a connection between influential users and misinformation, with one report showing that a relatively small number of Twitter users were responsible for spreading most of the misinformation in the 2020 US election (2021), and other finding a similar pattern on Facebook with anti-vaccine misinformation (Dwoskin 2021). James represents a different sort of influencer from the micro-celebrity commonly imagined: intentionally anonymous, followed directly by very few users, rarely engaging in two-way interaction, and yet having a significant cumulative effect on discourse regarding COVID-19 regulations. These users engage in an activity that we term to be firestarting: behaviors that generate high levels of antagonism and argument in social media spaces, but that rarely entail follow-up participation on those discussions. We identify firestarting as a subset of trolling, behaviors which are considered socially unacceptable within a community, and particularly in online communities (Cheng et al. 2017).

In this workshop paper, we apply a quantitative and qualitative analysis of this single user's impact on a dataset of Twitter reply conversations related to COVID-19. We suggest that a different approach to designing for discourse on social media is needed. Social media designers aiming to facilitate productive discourse and sensemaking on social media often suggest mass interventions, which aim to affect most or every user of a platform and cumulatively result in a superior discourse environment. Taking inspiration from Freed et al.'s work on intimate partner violence and technology usage, we suggest that designing for abusability, in which usability is specifically hampered for certain users, may be a profitable framework for those attempting to improve online discourse and minimize misinformation (2018).

Background

Current literature about online influencers primarily focus on their utility in marketing, and have particularly focused on measuring their profitability, attractiveness, trustworthiness, relationships, and other virality factors associated with their online presences (Chikhaoui, Chiazzaro, and Wang

2015; Woods 2016; Lou and Yuan 2019). More recent work has investigated social media influencers role in spreading misinformation. During the 2020 US election, researchers identified a core group of highly popular influencers who "actively promote and spread each others' content" about election-related misinformation (2021). The behavior of the subject of this study is notably different from these influencers, however, due to their lack of popularity as measured by followers and their lack of apparent connection to fellow influencers. This form of influencing also resembles trolling behavior, where users persistently and antagonistically violate community norms in online spaces (Tsantarliotis, Pitoura, and Tsaparas 2016). Cheng et al. describe how trolling differs from being an influencer in that the latter is a behavior, not an identity; anyone can potentially engage in trolling behavior, rather than a limited set of trolls (2017).

One design method for addressing antisocial behaviors is abusability. Abusability design and testing are critical perspectives, influenced by human-centered design and value-sensitive design, for technology development teams to consider potential risks that technology features may have for vulnerable communities (Chi 2020). Abusability centers sociotechnical consequences on human social, physical, and psychological harms, and acknowledges that usability may need to be decreased for some users in order to facilitate usability for others. Previous attempts to identify and moderate the ability of abusive users to influence platforms include Tweety Holmes, a tool for identifying abusive Twitter profiles (Kwon et al. 2018).

Data

We used Twitter's Streaming API to capture all tweets and replies made to 73 accounts associated with US governors, as well as the account for the mayor of the District of Columbia (DC). There are more than 51 accounts included because some governors operate multiple Twitter accounts. At least one account was captured for each governor. Data collection began March 28, 2020 and continued until November 1, 2020. A total of 12,544,760 tweets were collected in this time period. Governors' accounts in this period most often posted about the COVID-19 pandemic, including announcements of sometimes controversial statewide regulations, and also posted about the 2020 Black Lives Matter protests and other issues relating to state governance.

Within this sample, we pull all tweets posted in reply to these governors by James, for a total of 5,328 tweets. These include posts replying directly to the governor, and posts replying to other users who had replied to the governor. In addition, we had pulled a sample of James' timeline from May 16, 2020 to July 8, 2020, for a total of 3,864 tweets. Because James was suspended from Twitter and consequently had his tweets scrubbed from the platform and API, we have little insight into their Twitter behavior before the pandemic. However, Twitter's search function allows us to find the earliest time another non-suspended user had replied to them. Only one non-suspended user had ever replied to James before March 15, 2020, suggesting that before the COVID-19 pandemic they had received almost no attention on Twitter from other users.

Findings

Quantitative Analysis Out of a total dataset of 1,696,926 users who replied to our list of governor's accounts, James's conversations initiated the most follow-up replies at 28,085, more than twice the replies generated than the next ranked user on this statistic. Of all users analyzed, 67% had no follow-up replies to their posts, 1% of users had more than 100 follow-up replies, and only 337 users had more than 1,000 follow-up replies. James posted 5,328 replies (4th most in this dataset) for an average of 5.3 follow-up replies per governor's post (2nd most among users that posted at least 1,000 times). This combination of high post volume and high efficiency was unmatched in our dataset, with only one other user in the top 10 of both statistics (Figure 1).

James was unique in our dataset for having very few followers relative to their high level of follow-up reply generation. The median number of followers for users that generated more than 1,000 follow-up replies was 1,421, while James had 371 followers at their highest and under 100 followers for several months. They also stood in distinction to popular Twitter influencers with 10,000+ followers, who occupied 7 of the 10 spots for follow-up reply generation. These Twitter influencers would often make very few posts, and presumably relied on their popularity to attract their followers to their posts, a relative advantage James could not possess. James was also unique in our dataset for being the only user to post replies to governors' from all 50 states, with only 15 users posted in more than 40 states. Perhaps most notably, James posted incredibly quickly, with a median post time of under 10 minutes from a governor's original post, 4th among all users with more 1,000 replies to governors.

In our dataset of 3,864 of posts from James' timeline, 86% of their tweets were replies, 13% were retweets, and less than 1% of their posts were non-replies. During this period, the top six accounts they replied to were the governors of New York, Washington, California, Pennsylvania, Oregon, and New Jersey, although they also sometimes replied to non-governor politicians and popular right-leaning political commentators. Their top 10 retweeted accounts are all primarily political, pro-Trump accounts, with former United States president Donald Trump's account being the most retweeted. The content of James posts and retweets suggest engagement with the misogynist subculture of the "manosphere," as well as anti-Chinese government and anti-antifa rhetoric (Ging 2019). James posted every day of the week with reduced activity on Sundays, and posted every hour except for a five hour period from 11pm to 4am Pacific Time. This posting pattern closely aligned with other users' in our dataset. We stress, however, that extraordinary Twitter users may be extraordinary in their non-Twitter activities, and this user could well be operated by multiple people from different locations, or from a user with an unusual sleeping schedule located anywhere in the world.

Qualitative Analysis The content of James's tweets, the content within follow-up replies from engaged antagonistic Twitter users, and James's own follow-up responses. Particularly notable was James' tactic and posting almost the exact

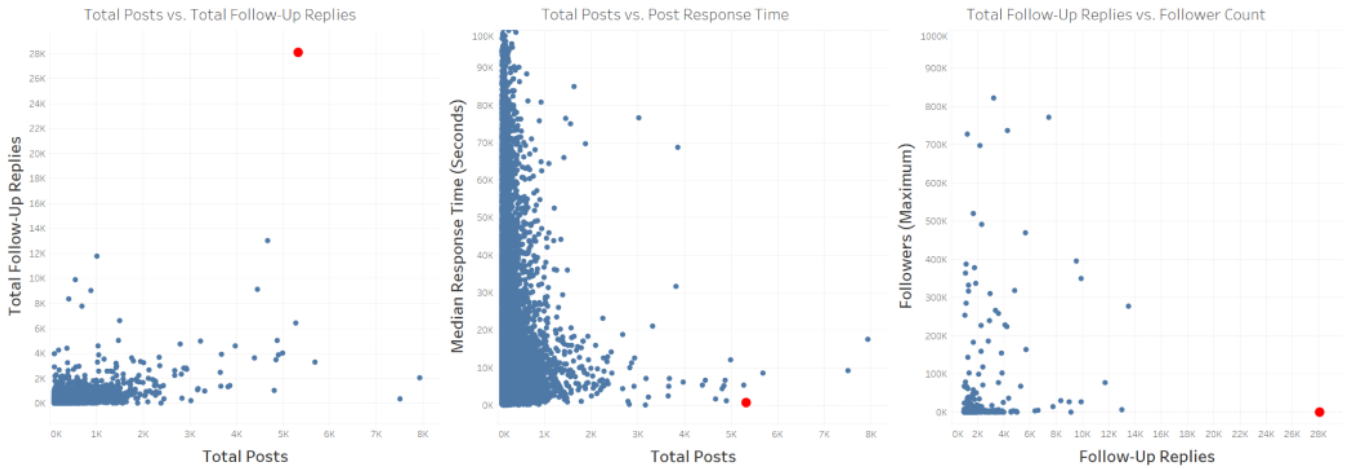

Figure 1: Three plots showing (from left to right) A) post volume against follow-up replies, B) post volumes against the median number of seconds a user posts' after a governor and C) follow-up replies (min 1,000) against a user's maximum followers. In each of these plots, the red dot identifies data associated with James, the central data point of this study. Extreme outliers have been removed for post response time and followers.

same reply to every state governor with only slight modification to localize each reply's content. For example, 1,146 of James's posts (29.6%) closely mirrored the same statistics and sentence structure as the tweets shown below.

> *James [2020-04-21]: @GovernorGordon Look at this data: Wyoming has 2 deaths out of 578,000 people. @GovernorGordon needs to open Wyoming for BUSINESS on Wednesday, April 22.*

> *James [2020-04-27]: @GavinNewsom California has 1,724 deaths out of 39,510,000 people. @GavinNewsom needs to open California for BUSINESS on Tuesday, April 28.*

James's claims did not include attribution to the source of the death rate statistic, although they tended to be accurate at the time of posting. In a similarly-structured tweet to New Mexico Governor Lujan Grisham, James's post engaged 154 follow-up replies, and in another tweet to California Governor Newsom, James's post received 666 follow-up replies. James's tweets in this format provoked similar engagement across different states, adjusted for that state's overall posting volume. This tactic, while taking a minimal amount of effort on James's part to look up death statistics, resulted in a high impact on other users' posting behaviors. Supporters of the governor tried to debunk James's statistical claims by including links to external articles about scientific studies validating high infection rates. This can be seen in James's second largest thread (717 follow-up replies), centered around Pennsylvania Governor Wolf. One user replies to one of James's statistical claims about the COVID-19 death rate:

> *@User2 [2020-05-12]: @James @GovernorTomWolf These are numbers from a real study. [external URL]*

James does not typically respond to these types of follow-up replies. Similarly, James does not engage with the many follow-up replies that claim the account is automated:

> *@User3 [2020-05-12]: @James @GovernorTomWolf Hold on.....this is essentially the same guy, with a hat in the profile, and replies to WA governor's threads.....I smell a bot.*

James does not need to defend themself from such claims, as other opponents of the governor will readily step in to engage their opponents. For example, supporters of James's claims would debate against opposers (such as User2, User3 above) by comparing the detriment of the economy to the deadliness of COVID-19:

> *@User4 [2020-05-12]: @UserA @James @GovernorTomWolf People can't live if there's no economy to return back to. There's massive food shortages, inflated prices, people will die of starvation. An economic collapse will lead to more deaths than the virus. Sit down, look at the bigger picture, and map it out.*

Other James supporters blame the governor for instilling public fear, and that "true" data exists to prove COVID-19 is not deadly:

> *@User5 [2020-05-11]: @UserB @James @GovernorTomWolf Stop fear mongering. Facts and data now indicate the virus is nowhere as deadly as we initially thought. Obviously, protect elders and those w/pre-existing conditions. Allow the rest of us back to work, we need to feed our families w/o government assistance!*

When James did engage with follow-up replies—a total of only 12 subsequent replies from James within the 717 follow-up replies—their statement would usually be a short, curt statement, occasionally with a link to an external news article to validate their original claims about misleading COVID-19 trends. In this case, in 11 of the 12 tweet replies, James reiterated their demand to reopen the economy and then linked to an article that claims that shelter-in-place orders were ineffective for preventing COVID-19 spread.

James [2020-05-11]: @UserC @GovernorTomWolf needs to drop Stay at Home orders. [external URL]

James will sometimes reply with non-sequitur posts, which often contain anti-China rhetoric, misogynist insults, references to the Christian bible, and claims that people become sick from COVID-19 because of unhealthy diets. For example, one often-repeated follow-up posted in nearly every state reads:

James [2020-05-12]: @UserD @GovernorTomWolf Tens of thousands of Pennsylvanians will die from poverty. More than the Communist Virus ever will.

James also strategically incorporates local political and social controversies only tangentially related to COVID-19 into COVID-19 threads, including in some posts references to local sports teams. In one example, James flooded California Governor Newsom's thread by framing the environmental plastic bag ban efforts as a dangerous tool in transmitting COVID-19:

James [2020-04-10]: @GavinNewsom when will you allow ALL Californians to use Plastic Grocery Bags, instead of the dangerous E-Coli Coronavirus infested cloth bags?

James will also craft new, particularly antagonistic tweets in response to current events. For example, the following tweet was posted in one of the first days of the 2020 Black Lives Matter protests:

James [2020-05-30]: @NYGovCuomo Black people eat more Sugar than others with different skin pigmentations. Sugar promotes Underlying Conditions that the Communist Virus targets. Any questions? [external URL]

This tweet is an exemplar of James' posts: non-sequiturs related to hot-button health issues like sugar consumption, flagrantly racist and aligning with pro-Trump phrases like "communist virus," and most importantly, easily falsifiable by a regular commenter on Twitter. These phrases provoke enormous response both from those who perceive themselves as having a commitment to truth on the internet, and from those whose political ideologies (racist, anti-China) align them with the content of James' tweets regardless of their veracity.

Discussion

In this paper, we describe a user whose systematic, strategic, and antagonistic behavior has ensnared thousands of Twitter users in bad faith, bitter arguments during a prolonged crisis. This user, who we specifically refer to as James and in general refer to as a firestarting troll, combines a time-intensive method of replying to every United States governors' posts with a set of messaging tactics seemingly designed to encourage follow-up replies from their fellow users. These messaging tactics exploited their fellow users' apparent desire to correct misleading information, respond to locally-relevant controversies, and fight back against bigoted rhetoric, and their rapid and uninterrupted posting schedule likely allowed them to exploit Twitter's reply-ranking algorithm. They achieved this level of disruption despite having extremely few Twitter followers and little follow-up engagement with their initial posts.

James is doubtless a curiosity, but we argue that their behavior also has implications for platform moderation. One account had an significant and outsized effect on Twitter public discourse, sowing misinformation, misogynist and racist views, and non sequitur arguments that draw in thousands of other users replying to governors' of every state. James abused the design of Twitter's threaded discourse system to disastrous effect, likely using automated or semi-automated posting methods to reply quickly to governors' posts and boost their visibility in Twitter's reply-ranking algorithm. Rather than focusing on design for the great majority of users' behaviors, designers of discourse systems may find it profitable to focus on designing for the minority of users that have an outsize and negative effect on discourse results. In this case, monitoring user-level statistics on how and how often they engage in reply conversations, and provisioning additional scrutiny and moderation actions such as temporary or permanent suspensions, would have easily identified and mitigated James' particular method of posting. While some may argue that if James is suspended, another user would simply take their place, we note that James' methods and tactics required a level of nonstop dedication, planning, and cultural knowledge unlikely to be replicated by their peers.

This analysis of one user cannot capture the full scope of firestarting behaviors on Twitter's platform, a clear limitation of this study that will be addressed in future studies on the full scope of firestarting users in this dataset. The uniqueness of James's account is also highly contingent on our choice of dataset; for example, there may be an even more prolific firestarter that specifically responds to United States mayors or science journalists, rather than governors. However, we note that James' decision to target governors' was likely tactical, given the valuable information they provide to a broad and interested public during crisis events. Moderators during crisis situations may do well to actively monitor high-importance and high-traffic Twitter accounts, like those of governors, to proactively identify firestarter accounts like James.

Acknowledgements

This research was supported by the National Science Foundation COVID-19 Rapid Response Research (NSF RAPID) program (2027792). We also wish to thank UW Center for an Informed Public for providing infrastructure support.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, 65–74. New York, NY, USA: Association for Computing Machinery.

Center for an Informed Public; Lab, D. F. R.; Graphika; and Obervatory, S. I. 2021. The Long Fuse: Misinformation and the 2020 Election. Technical report.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1217–1230. New York, NY, USA: Association for Computing Machinery.

Chi, N. 2020. What is Abusability Testing and Why is it Necessary? | Hacker Noon.

Chikhaoui, B.; Chiazzaro, M.; and Wang, S. 2015. Discovering and tracking influencer-influencee relationships between online communities. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 1–9.

Dwoskin, E. 2021. Massive Facebook study on users' doubt in vaccines finds a small group appears to play a big role in pushing the skepticism. *Washington Post*.

Freed, D.; Palmer, J.; Minchala, D.; Levy, K.; Ristenpart, T.; and Dell, N. 2018. A Stalker's Paradise: How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery. 1–13.

Ging, D. 2019. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities* 22(4):638–657. Publisher: SAGE Publications Inc.

Kwon, S.; Liang, P.; Tandon, S.; Berman, J.; Chang, P.-j.; and Gilbert, E. 2018. Tweety Holmes: A Browser Extension for Abusive Twitter Profile Detection. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18, 17–20. New York, NY, USA: Association for Computing Machinery.

Lou, C., and Yuan, S. 2019. Influencer Marketing: How Message Value and Credibility Affect Consumer Trust of Branded Content on Social Media. *Journal of Interactive Advertising* 19(1):58–73. Publisher: Routledge _eprint: https://doi.org/10.1080/15252019.2018.1533501.

Tsantarliotis, P.; Pitoura, E.; and Tsaparas, P. 2016. Troll vulnerability in online social networks. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, 1394–1396. Davis, California: IEEE Press.

Woods, S. 2016. #Sponsored: The Emergence of Influencer Marketing. *Chancellor's Honors Program Projects*.