Reporting Hate Speech and Disinformation on Twitter

USER-EXPERIENCE ANALYSIS AND RECOMMENDATIONS

JESSICA YOUNG

INTRODUCTION

This paper assesses current protocol for reporting abusive social media content, with an in-depth focus on Twitter. It follows a larger body of research dedicated to detecting common threads across instances where abusive content that spread online caused inordinate physical harm offline. One problematic and reoccurring theme found in "Beyond AI: Responses to Hate Speech and Disinformation," was a cumbersome and confusing user experience for stakeholders when reporting content found to be at the root of harm. In the US, India, and Myanmar, individuals receiving threats, NGOs witnessing threats, and law enforcement officials responding to threats repeatedly echoed this concern after trying unsuccessfully to have abusive posts removed.[i] Gleaning lessons from these case studies, this paper (1) acknowledges current best practices for soliciting abusive content reports; and (2) makes recommendations to alleviate current points of friction and to seize opportunities to collect data pertaining to known common threads. These recommendations prioritize the detection and removal of content likely to cause the greatest (physical) harm.

USER-EXPERIENCE BENCHMARKING & ANALYSIS

Considering all of the platforms reviewed in our research (e.g. Facebook, WhatsApp, Reddit and Twitter), Twitter has the most thorough interface for reporting harmful content.[ii] The user experience Twitter offers is thorough in three key regards: the Help Center it provides for users, the granularity of reporting available, and the breadth of information collected on each case reported.

Given that content moderation policies vary from platform to platform, we consider a Help Center for reporting violations to be a best practice. When in doubt *during* the reporting process, Twitter users can click through to a simple guide with minimal, scannable text. This guide provides an overview of how to report content (whether a tweet or direct message) that violates Twitter rules (community standards) and/or terms of service. Recognizing that media outlets (e.g. Infowars, as covered in our paper) and individual profiles can be repeat offenders, it also provides how-to information for reporting media and profiles. Those unfamiliar with what qualifies as a violation on Twitter may explore a side menu that clearly lays out Twitter rules and policies, general guidelines, as well as law enforcement guidelines[iii].

When right-clicking above a Tweet, a user immediately sees an option to launch a report. In order to classify the type of problem at hand, the user is asked to indicate the issue – from merely suspicious content to clearly abusive or harmful content. A second screen seeks valuable granularity to decipher both the context of the problem and its severity. Options for flagging include the following: "disrespectful or offensive" (Appendix A); "private information"; "targeted harassment" (Appendix B); "directs hate against a protected category" (Appendix C); "threatening violence or physical harm" (Appendix D); or "encouraging or contemplating suicide or self-harm."[1] Given the common threads noted in our case studies[iv], having categories for directly reporting both hate speech and content eliciting harm should be considered a best practice. We note that hate speech, should it not fall into one of the protected categories listed, could fall into the category of targeted harassment. In cases of harassment, hate speech against a protected category, or threatening violence or physical harm, Twitter again collects more context on whether such content was directed against the user reporting the violation, someone else, or a group of people. For the hate speech and violence/physical harm categories, Twitter capitalizes upon user insight by allowing up to five abusive tweets from the account-holder to be attached to the report (Appendix C). We consider this an excellent means of adding weight to a report and potentially escalating its review; as noted in our research, in the most egregious cases explored, accounts eliciting the most harm were often repeat offenders. This consistently held true whether they were humans (e.g. military officials in Myanmar) or bots (e.g. in the case of Pizzagate). Depending on the type of content flagged, the user filing the report sees three to four screens total before exiting the relatively quick but thorough process. The user-experience concludes with a 'thank you' and an additional best practice: letting the user know what will happen next. Twitter clarifies that if the content is found to be in violation of its

[1] An appendix highlights the reporting process for the four categories relevant to the case studies we have explored. For instance, content referenced in Myanmar containing ethnic slurs and calling for murder could be at once considered offensive, hate speech against a protected category, and threatening violence or physical harm. Pizzagate posts on the other hand did not contain hate speech against a protected category but were nonetheless harassing and hateful in nature.

policies, it will be removed (Appendix A). While a simple step, this is lacking on other platforms. While we think additional details should disclose a turnaround time, we find the absence of even this basic statement to be problematic; civil society organizations repeatedly complained that when flagging content, no clear information was provided on next steps which led them to repeatedly flag the same harmful content when it remained online[v].

In sum, this interface implements several best practices to identify problematic content and to smoothly engage users in the process. (1) It allows contextualized reporting with follow-up screens customized according to the abusive content at hand. (2) It seeks to directly identify cases likely to cause harm and classify them based on the degree of harm (e.g. violence or physical harm). (3) It seeks to identify repeat offenders frequently posting abusive content. (4) It keeps the user in-the-know to provide the best report possible (e.g. through a Help Center, back and forth navigation between screens, and a 'thank you' screen letting the user know what to expect next).

RECOMMENDATIONS TO ESCALATE ABUSIVE CONTENT REMOVAL

While Twitter's interface for reporting harmful content includes several ideal elements, we find additional areas for improvement based on the common threads surfaced in our cases studies. That is, many posts that contributed to rampant physical harm contained elements that would go un-noted in Twitter's reports as they stand. We note five main areas for improvement in the reporting interface itself, as well as one recommendation for the processing of each report once filed.

First, we see a significant opportunity to collect context around where the issue resides (e.g. in the text of the post shared, in the post's media such as a photo or video, or in the hashtag accompanying the post). We noted in our cases studies, non-governmental organizations (NGOs) complained that they would flag abusive content only to have it remain when content moderators failed to look beyond the post's text.[vi] Further, we saw a common thread of platform users attempting to game the content moderation policies.[vii] Such users often embed abusive content in photos or videos – a problematic practice given AI is less capable of detecting abusive keywords in such media. In the case of Twitter posts, we note that users may also game the system by stringing together abusive content in a hashtag or a string of hashtags rather than in the post's text itself (Appendix E); this again, would be more difficult for AI to parse than distinct words. Given the above, we recommend an option to select where the problematic component is found: in the post's text, image, video or hashtag. (While it shouldn't be mandatory, if the user is able to note a timecode, this would also be worthwhile to avoid overlooking issues in lengthy videos.)

Second, we recommend allowing users to select multiple options when applicable. Given especially inflammatory posts may contain all (or nearly all) of the categories available[viii] [ix], it can be ambiguous for the user to know which option should be selected. In the instance where multiple classifications apply, it is also unclear which category is considered the most severe; if this was apparent or signaled to the user, they would be better equipped to select the most appropriate category for their report. When assessing common threads across our case studies, we found abusive content that resulted in the most extreme physical harm offline tended to combine *both* hate speech and threats of/calls for violence. Thus, we strongly recommend there be an option to flag when content includes both of these factors. Such reports should then be escalated given past trends.

Third, while a matter of semantics, reports of content involving violence should be broadened in scope. In our research, we note abusive content found to cause harm not only included threats of violence (a current reporting option in Twitter) but also broader calls to violence.[x] The fact that intent to harm in the case of Pizzagate and actualized harm in the case of Myanmar both stemmed from posts eliciting harm provide support for this argument.

Fourth, there should be an additional option to flag disinformation. Here, we note that unlike other platforms (such as Facebook[xi]), Twitter has intentionally left this out of their reporting process. If a company wishes to avoid policing disinformation on its platform, it should at the very least permit flagging disinformation as an associated factor when one selects the hate speech or violence/physical harm categories. This too is influenced by our finding that disinformation is often paired with hate speech or calls to violence to amplify the degree of offline harm[xii].

Fifth, as a best practice, we recommend the final screen of a report contain an expected processing time. This should (a) provide a measure of accountability and transparency for enforcing terms of service and community standards, and (b) provide users a timeframe for following up. In the case of Myanmar, we saw NGOs took it upon themselves to constantly check if violence-inducing content had indeed been removed after it was flagged – with platform processing speeds far exceeding those publicized.[xiii] To mitigate the violence that erupted, such organizations would repeatedly flag the same

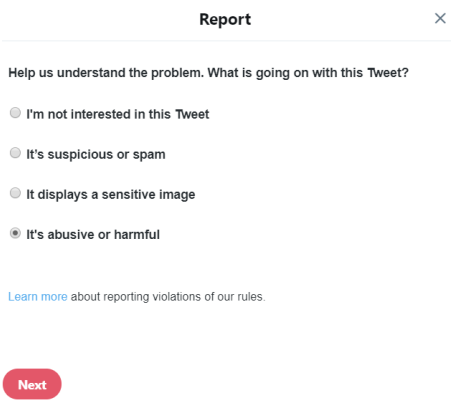
content and contact platform representatives directly until the issue was addressed. To alleviate this time-intensive process and the burden placed on vigilant users, platforms should provide a timeframe for processing one's complaint.

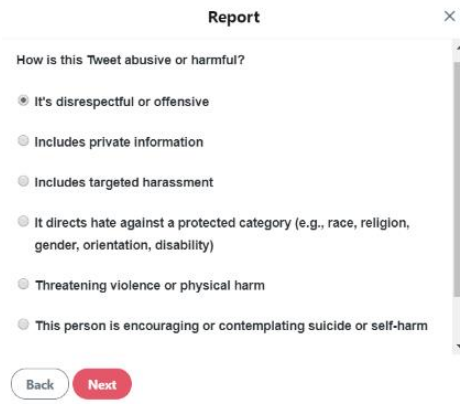
Finally, while not a part of the user interface, we recommend a back-end feature to assist with report processing. Drawing from other common threads observed as to what types of accounts have historically realized the greatest reach for their abusive content, we recommend that systems on the back-end assign a bot-probability score and note the number of followers for an account being flagged[xiv]. These variables should be used along with the abuse categories flagged when assigning weight to a report. Those reports containing a combination of the most problematic features for widespread harm (e.g. high probability of being a bot, high follower count, and threats of violence combined with disinformation and/or hate speech) should receive escalated priority.

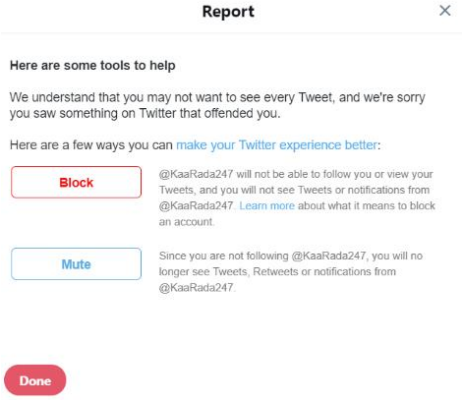
APPENDIX

Twitter 'Report Tweet' Interface for Abusive or Harmful Content[xv]

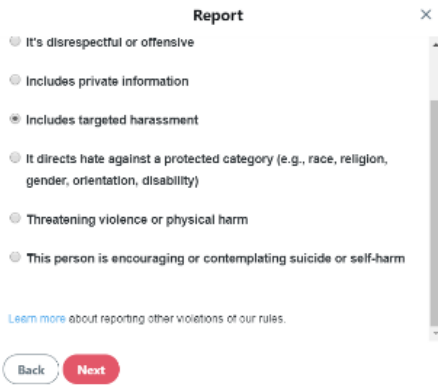
A. Disrespectful/Offensive Content

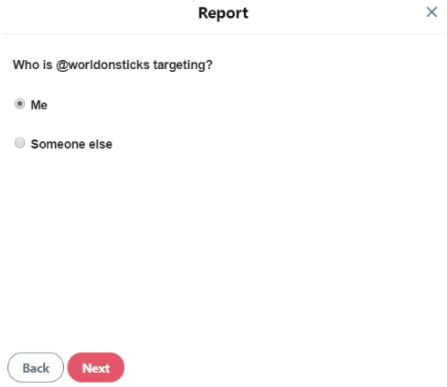




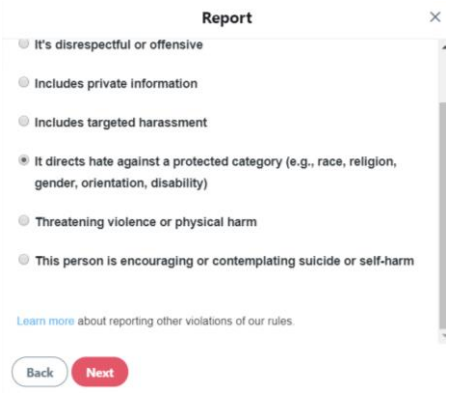


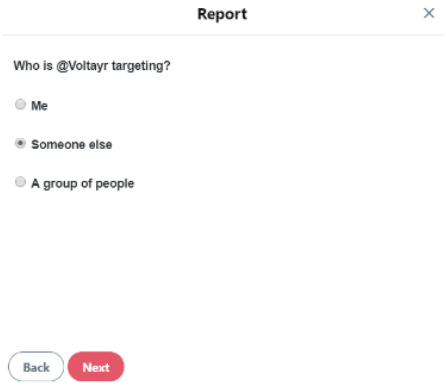
B. Targeted Harassment

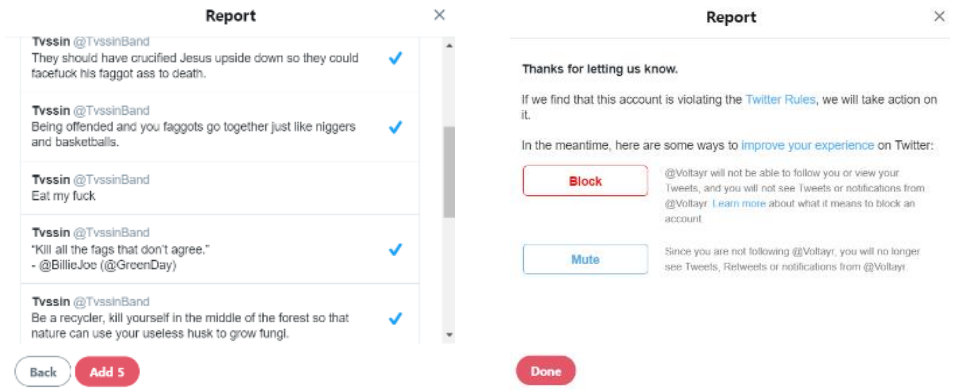




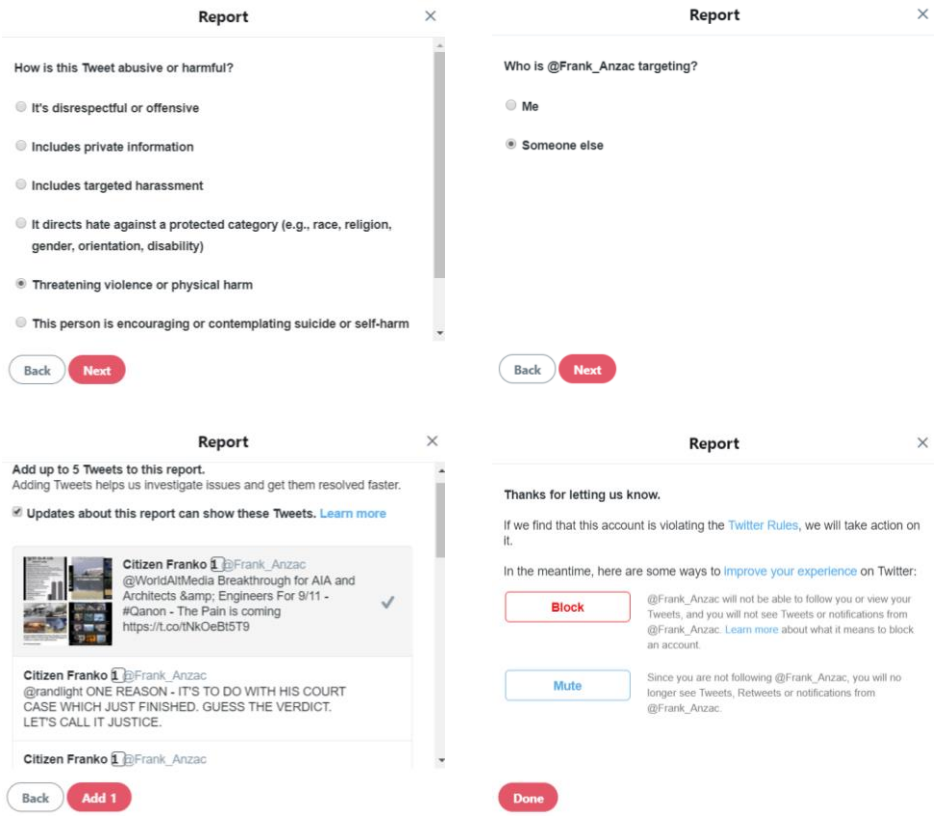
C. Hate Speech Against a Protected Category







D. Threatening Violence/Physical Harm



E. Gaming Content Moderation Systems

[i] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[ii] Twitter. 2018. Report Violations Page. *Twitter*, Accessed December 15, 2018. https://help.twitter.com/en/rules-and-policies/twitter-report-violation#specific-violations.

[iii] Ibid.

[iv] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[v] Rajagopalan, M., Vo, L.T., and Soe, A.N. 2018. How Facebook Failed The Rohingya In Myanmar. *BuzzFeed News*, Aug. 27, 2018.

[vi] Stecklow, S. 2018. Why Facebook Is Losing the War on Hate Speech in Myanmar. *Reuters*, August 15, 2018.

[vii] Rajagopalan, M., Vo, L.T., and Soe, A.N. 2018. How Facebook Failed The Rohingya In Myanmar. *BuzzFeed News*, Aug. 27, 2018.

[viii] Specia, M., and Mozur, P. 2017. A War of Words Puts Facebook at the Center of Myanmar's Rohingya Crisis. *New York Times*, October 27, 2017.

[ix] Safi, M. 2018. Revealed: Facebook Hate Speech Exploded in Myanmar during Rohingya Crisis. *The Guardian*, April 3, 2018.

[x] Stecklow, S. 2018. Why Facebook Is Losing the War on Hate Speech in Myanmar. *Reuters*, August 15, 2018.

[xi] Facebook. 2018. How Do I Mark a Post as False News? *Facebook*, Accessed December 15, 2018. https://www.facebook.com/help/572838089565953

[xii] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[xiii] Rajagopalan, M., Vo, L.T., and Soe, A.N. 2018. How Facebook Failed The Rohingya In Myanmar. *BuzzFeed News*, Aug. 27, 2018.

[xiv] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[xv] Twitter. 2018. Report Tweet Interface. *Twitter*. Accessed November 19, 2018. https://twitter.com/.