Introduction

This literature review examines articles written about harassment on social media microblogging platforms. Since there's a large body of writing about (pre-Musk) Twitter, it draws primarily on those sources, but many of the lessons learned there can be applied to the Fediverse as well. The purpose of the literature review is to examine previous patterns of harassment on Twitter and the Fediverse. The literature is largely written by Black women, and reflects how they have used the available tools the software and community offer them to tackle the abuse. It also demonstrates how the Fediverse has had similar issues of harassment and what was done there to battle said harassment.

Intersectionality as an analytical framework is relevant in that much of the literature examines how Black women and other marginalized users are the main victims of online harassment in social media, but also because Black women have tended to take the lead with regards to documenting online harassment and placing it within its proper sociohistorical context. We refer to intersectionality in a two-fold way here, then: it is a theoretical/epistemological lens that informs this literature review, and (on a more pragmatic level) it is also a taxonomic tool for moderators and admins which can help them correctly identify harassment, and learn how to understand problematic behavior and why certain users may feel as if they are the target of harassment.

As part of a series of articles on the topic of anti-harassment features, the goal of this first article is to establish the current state of play by enumerating common harassment vectors. Subsequent articles, by contrast, will focus on alleviating harassment by recommending design considerations to close off some of the vectors described here. In other words, examination of the literature in this article will demonstrate the more prevalent forms of harassment being experienced, and subsequent work will provide ideas as to what social media in the future needs to create and utilize so that moderators and admins can provide safer environments for users.

Online Harassment Against Vulnerable Populations

In 2018 Amnesty UK conducted research about the abuse women face on Twitter. From the report:

> ... **#ToxicTwitter** - the result of interviews with more than 80 women, including politicians, journalists, and regular users across the UK and USA - Amnesty exposes how Twitter is failing to respect women's rights, and warns the social media company that it must take concrete steps to improve how it identifies, addresses and prevents violence and abuse against women on the platform.
>
> Amnesty UK

A particular highlight of the research Amnesty UK reported was abuse against public figures, specifically Black women who are Members of Parliament (MPs):

> Reports of horrifying levels of online abuse against black women MPs are deeply concerning.Our research has revealed the shocking levels of abusive tweets hurled against women of colour in politics and public life - especially black women, who were found to be 84% more likely than white women to be mentioned in abusive or problematic tweets.
>
> Amnesty UK

Hypervisibility experienced by Black women, be they regular users or MP's, creates a spotlight on their accounts and opens them up to untold violence, particularly if they discuss racism and sexism, and engage in the use of public campaign hashtags. As written in an article on hypervisibility for Bloomberg:

> Members of marginalized communities experience the distress of hypervisibility in the workplace and beyond—the feeling of being

> overly visible because of an individual's race or ethnicity, sometimes to the point of overshadowing their unique skills and personality.
>
> This type of extreme focus on skin colour in the workplace instead of the substance of the professional in the context of the relevant business can sometimes detract from due recognition, reward, and cultivation of employees' distinct talents as a technician, for example. It can also be detrimental to the mental health of individuals who experience this type of attention.
>
> Bloomberg

Intersectionality posits that being a woman and being Black means that the misogyny experienced is antiblack and colorist, and if the user is queer or disabled, the prejudice can take on these other forms as well. An intersectional perspective also accounts for the fact that the abuse of Black women can also be used as a vehicle for bad actors to create discord in other arenas not immediately or obviously related to Black womanhood. In other words, an action undertaken by either an individual, group, or company that is antagonistic towards Black women (other marginalized groups are also used in this way) is able to drive "clicks", and generate traffic to support the harasser's cause. This traffic has the capacity to impact wider culture, and bad actors have worked to see this kind of abuse come to fruition:

> There becomes a pseudo Overton Window lock: Harmful behavior toward Black women isn't enough to inspire change until others are harmed, but the original harms are often lost by journalists tasked with covering tech. The power and rhetoric that went unchecked becomes common. And the tactics used against Black women for "lulz" become weapons used in the conspiracies destabilizing the very nature of truth, from the swarming of victims to posing as

> Black women to destabilizing communities (or countries). Defining the systemic abuse becomes a frustrating exercise of describing an empty space that no one believes is there. If we can follow, surveil, and automate everyone, how could we miss anything important? And if it is important it is only important for how it changes the mythical “standard user” no matter how many are hurt before.
>
> Sydette Harry

As Harry shows, abuse of Black women on Twitter and elsewhere doesn’t “just” impact Black women; it can be used as part of a wider strategy to attack movements that those Black women belong to–or are seen as belonging to–adding insult to injury as Black women become cannon fodder for larger harassment campaigns.

While harassment aimed directly at Black users is bad enough, Sydette Harry also gestures towards a more insidious form of harassment: "posing as Black women" in order to "destabiliz[e] communities". Combatting this phenomenon--known as "sock puppeting" or "creating sock puppet accounts"--often requires intimate knowledge of the norms of the community being targeted. As Bret Schafer writes:

> The ability of foreign actors to look, act, and speak like the online communities they target creates the obvious potential for manipulation. At the same time, genuine activists may, in fact, be best positioned to recognize and ferret out the imposters in their midst. For example, in the aftermath of the killing of Philando Castille, an unarmed Black man shot to death by a police officer in Minnesota, Black Lives Matter activists flagged what turned out to be a Russian-operated, faux-BLM Facebook page as suspicious due to its use of the slogan “Don’t Shoot”—a phrase that many genuine activists had long since abandoned.

> Those subtle inconsistencies are less likely to be noticed by those outside of Black activist circles, meaning that content seeded by IRA trolls posing as Black activists may have a more profound impact on broader public perceptions and opinions. This is especially problematic given that Russian internet trolls often engage in a form of digital blackface, painting grossly stereotypical portrayals of the African Americans whose personas and vernacular they co-opt. These caricatures are not only offensive and demeaning, they can discredit the legitimacy of the causes they claim to support. By promoting viewpoints that seek to generate polarization rather than reasoned debate, Russian trolls can skew the perception of critics and potential supporters alike, particularly when those opinions—like Luisa Haynes'—find their way into mainstream media outlets.
>
> Bret Schafer

As Schafer shows, while it is fairly trivial for trolls to create sock puppet accounts imitating Black users, rooting out those accounts requires specialist knowledge and time. Unfortunately, between the moment that sock puppet accounts are created, and the moment they are discovered (if ever), they have the potential to skew discourse away from genuine issues and "generate polarization rather than reasoned debate", which can harm the perception of imitated communities by "mainstream media outlets". In other words, when their prejudices about Black folks are being played into by sock puppet accounts, social media users (and wider media) who would otherwise be "on the fence" about an issue can be radicalized in a more right-wing direction by the perception that one side is reasonable and calm, and the other is inflammatory and hostile. This is demonstrated through examples in the next section.

Examples

The different types of harassment enumerated about are exemplified by specific incidents such as GamerGate, the fake hashtag “#EndFathersDay” on Twitter, and similar incidents mirrored in the Fediverse, which placed Black femme users at the center of harassment. The articles written by Ra’il I’Nasah Crockett and Sydette Harry both discuss how these incidents demonstrate surveillance upon Black women, and how disposable Black women’s safety is, while also illustrating the efficacy of sock puppet accounts:

> Consider the #EndFathersDay hoax carried out by 4chan and the resulting Black feminist #YourSlipIsShowing counter-attack. 4chan’s attack was fundamentally on feminism itself, #EndFathersDay as a hashtag was meant to make feminism look like a form of anti-male extremism, turning moderates away from feminism while strengthening its opponents. While many mainstream media outlets swallowed the hoax hook and all, several individual Twitter users, led by @sassycrass and including myself, instead began to do some basic detective work. What we uncovered was an extended year-long plan, where 4chan users were to set up fake accounts where they would pretend to be Black women, women of color, trans women, and otherwise marginalized folks, infiltrate our spaces, study how we operate, then wage hashtag war.
>
> Ra’il I’Nasah Crockett

Large accounts held by women of color on the Fediverse in 2019 (and they were on a Black led instance) were also at the center of a large argument between other instances on the Fediverse, whose flames were fanned by a white male user who had created a hashtag and attached the names of the women of color; these women were not involved in the dispute between the other instances and had nothing to do with the

creation of the hashtag. But due to their hypervisibility and large followers, the tag stuck to them and they were then attacked and attached to the dispute.This is similar to what happened on Twitter:

> [The trolls] were successful, not just because they capitalized on the ever-present misogynoir within the mainstream feminist movement, but because stalking Black women online at this point is a common, acceptable practice. Of course this misogynoir-fueled stalking is usually done in the “name” of something, and that something is always implied as being better and greater than Black women.
>
> Ra’il I’Nasah Crockett

Due to Twitter being centralized, users were able to connect with each other quickly in order to mitigate some of the harm, through direct interaction and the hashtag #YourSlipIsShowing (a Southern American saying which suggests that what the subject wants to hide is inadvertently visible), which was used to sound the alarm on a troll or fake profile that other users would have been able to see and block immediately. Due to the decentralized nature of the Fediverse, by contrast, with its inevitable fragmentation of discourse, victims of abuse may not even realize they are being lined up in the sights, or by whom. In this case, hashtags (which are surfaced more readily than non-hashtagged messages by most softwares) are not just nice-to-have as anti-harassment tools, but can be absolutely vital for promoting awareness of nascent and ongoing harassment campaigns.

Conclusion

This article has recounted the complex landscape of online harassment within Twitter and the Fediverse, shedding light on various types of misconduct that threaten safe online digital environments throughout recent years. Here, the insidious nature of direct harassment was explored, where individuals are directly targeted with abusive content and threats. Also explored are the deceptive tactics employed in sock puppetry, where perpetrators hide behind multiple fake accounts to amplify their harmful actions. There was also a focus on antiblack misogyny perpetuated onto Black women in both Twitter and Fediverse, particularly through the use of fake hashtags posted by sock puppet accounts, and also how Black women used hashtags as a means of spotting bad actors and combating the harassment and abuse. Similar actions by bad actors occurred on the Fediverse, and so Fediverse users employed the same alarm bell tactics to warn others about these bad actors and hate campaigns.

From the outset, the promise was to approach these issues through an intersectional framework, acknowledging the unique experiences of marginalized communities. As we continue to explore the Fediverse's challenges and opportunities, it becomes apparent that harassment is often compounded by factors such as race, gender, sexuality, and disability. These intersecting identities can make certain individuals even more vulnerable to online abuse, necessitating a nuanced and inclusive approach to addressing these issues effectively.

Future articles within this series will further discuss these issues and also look at how Fediverse software has the ability to mitigate harm to various communities through the implementation of tools that are functional and easy to use.

References + Further Reading

Amnesty International. “Toxic Twitter: The Solution”. Amnesty, 2018.

Amnesty International UK. “UK: online abuse against black women MPs 'chilling'”. Amnesty, 2020.

Berger, Cathleen. “Mastodon as a common good alternative: Conditions apply”. Reframe Tech, 2022.

Bloomberg. “Recognizing and preventing the strain of hypervisibility”. Bloomberg, 2021.

Crockett, Ra'il I'Nasah. “‘Raving Amazons’: Antiblackness and Misogynoir in Social Media”. Model View Culture, 2014.

Eleanor. “Mastodon 3.0 A short overview of some of our newest features”. Mastodon Blog, 2019.

Farokhmanesh, Megan. “Twitter rival Mastodon isn’t safe from online mobs either”. The Verge, 2018.

Goldberg, Michelle. “Feminism’s Toxic Twitter Wars”. The Nation, 2014.

Hampton, Rachelle. “The Black Feminists Who Saw the Alt-Right Threat Coming”. Slate, 2019.

Harry, Sydette. “Listening to Black Women: The Innovation Tech Can't Figure Out”. Wired, 2021.

Lauren, Genie. “Twitter Stole From Me and They Can Steal From You, Too”. The Root, 2018.

O'Sullivan, Donie. “American media keeps falling for Russian trolls”. CNN Business, 2018.

Prokop, Andrew. “Justice Department charges Russian national with conspiring to interfere with 2018 midterms”. Vox, 2018.

Rochko, Eugen. “Learning from Twitter’s mistakes: Privacy and abuse-handling tools in Mastodon”. Mastodon Blog, 2017.

Rochko, Eugen. “Cage the Mastodon: An overview of features for dealing with abuse and harassment”. Mastodon Blog, 2018.

Schafer, Bret. “Race, Lies and Social Media: How Russia Manipulated Race in America and Interfered in the 2016 Elections”. State of Black America, 2019.