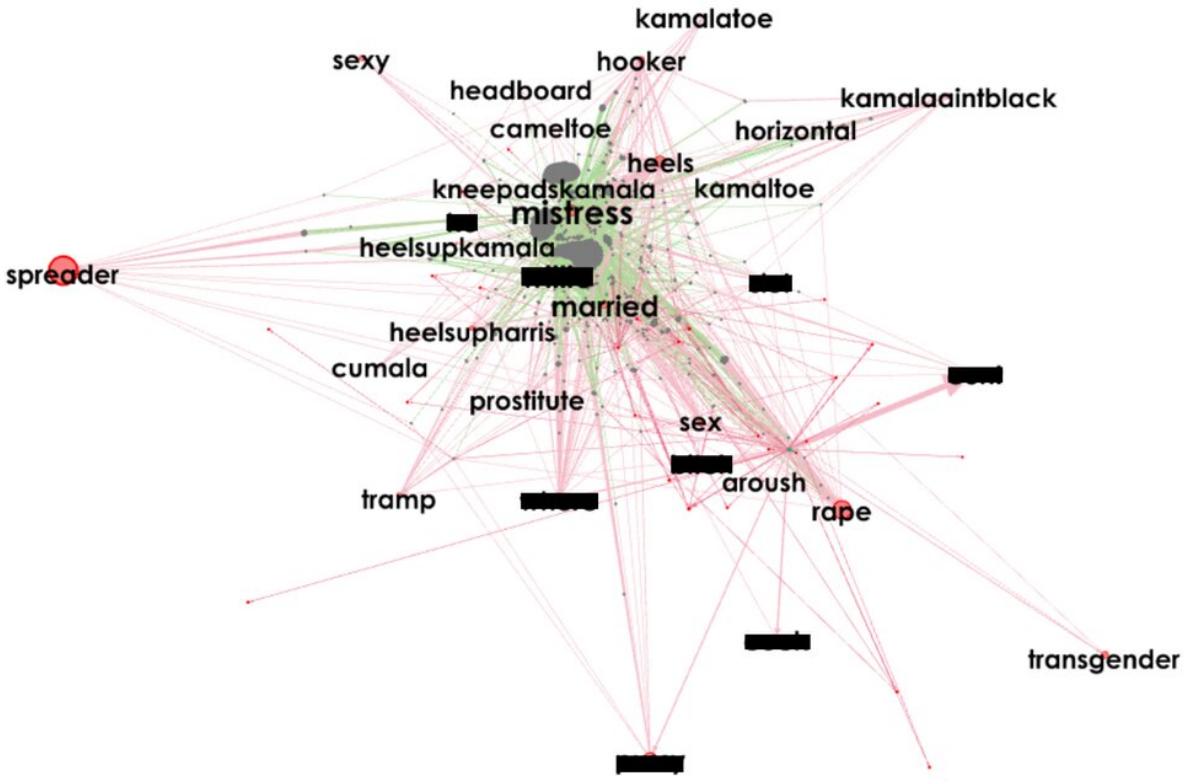
Kamala Harris abuse campaign shows how trolls evade social media moderation

January 26 2021, by Alexandra Pavliuc



A network visualisation showing some of the abusive terms, often coded, shared about Kamala Harris between September and November 2020. Credit: Wilson Center Malign Creativity Report

As Vice President Kamala Harris settles into her first full week in the White House, thousands are heading online to celebrate her

groundbreaking achievement. Unfortunately, thousands more are flooding social media with sexualised, transphobic, and racist posts which continue to highlight the particular abuse faced by female politicians online.

My colleagues and I studied this abuse in the weeks leading up to the 2020 US presidential election, revealing how trolls and abusers commonly use coded language and dog whistles to evade the moderation efforts of social media companies. This evolution of online hate speech undermines the "automated moderation" tools that platforms are currently using to tackle hate speech.

As well as abuse and hate speech, our team tracked the proliferation of "gendered and sexualised disinformation narratives" about 13 female politicians across six social media platforms. These false or misleading narratives are based on women's gender or sexuality, and are often spread with some degree of coordination.

In one example of gendered disinformation, a former Ukrainian MP was targeted with doctored images of her running naked around the streets of Kyiv on Twitter. The former MP has stated that the image and narrative still circulates online whenever she does public work abroad.

Evading moderation

Our study, which collected data between September and November 2020, found over 336,000 abusive posts on Twitter, Reddit, Gab, Parler, 4chan and 8kun, 78% of which targeted Kamala Harris. That equates to four abusive posts a minute over the course of our two-month study period—and three a minute directed at the woman who is now vice president of the United States. Each of these posts had the capacity to be shared thousands of times, exponentially multiplying their reach.

Social media companies are using automated content moderation tools in an effort to flag and delete gendered hate speech quicker. These tools are designed to instantly detect harmful social media posts, but they can only do so by being told which words are considered abusive. This leaves a "blind spot" for any abusive language which has yet to be flagged as abusive by human moderators.

Our research shows that online abusers are evolving the langauge they use in order to avoid detection by moderation tools. We call this process "malign creativity," which we believe to be a significant challenge social media companies must overcome if they are to conduct effective content moderation at scale.

We've observed abusers crafting false narratives and memes, tailored to the female politician they seek to harass, and shrouded in coded language. An example of a false sexualised narrative we saw against Kamala Harris was that she "slept her way to the top" and is therefore unfit to hold office. This narrative spread across platforms with hashtags that no automated classifier could detect without being pre-coded to do so, such as #HeelsUpHarris or #KneePadsKamala.

Through network visualization, we saw that some users who engaged with this narrative also engaged with other sexualised, racist and transphobic narratives. This underscored the intersectional abuse that women of color face online.

Prioritizing moderation

Coded language and dog whistles (which are subtle messages designed to be understood by a certain audience without being explicit) make detecting gendered and sexualised disinformation on social media particularly difficult without high levels of investment in detection technology. That's why our report recommends social media platforms

update their content moderation tools to pick up on new and emerging narratives that demean the world's most powerful women.

This should be done in coordination with the women themselves, or their campaign and marketing teams. Platforms must also allow women to submit "incident reports" that cover multiple individual posts, rather than forcing them to report each piece of abusive content, one at a time—which is both laborious and upsetting.

Gendered and sexualised disinformation affects the public's perceptions of high-profile women. Some women at the beginning of their careers may feel harder hit by gendered disinformation and abuse, choosing not to enter public-facing careers at all due to the abuse they see targeted at others—and at themselves.

One recent study found that women decrease their engagement online to avoid ongoing or potential harassment. One of those interviewed in the study resented that she had to "wade through all this filth… to just do the basic function of participating" on social media. These sentiments highlight the harsh reality that women must accept if they wish to engage online, where harassment is more sustained and violent compared to what men face.

Kamala Harris' historic inauguration has been cause for celebration for women, and women of color especially. It's also an opportunity for a new generation of women to feel inspired to pursue leadership roles.

To ensure women are inspired by the presence of other women in high office and not dissuaded by the abuse they may face, social media platforms and governments are responsible for providing spaces in which women can participate equally online. Effectively moderating the gendered abuse women suffer on social media—especially that which passes undetected by automated tools—is a crucial part of that

responsibility.