IJHSR
ijhighschoolresearch.org

■ RESEARCH ARTICLE

Gendered Online Toxicity During COVID-19

Christine Lee

Riverdale Country School, 5250 Fieldston Rd, The Bronx, NY, 10471, USA; christine.lee.nyc8@icloud.com
Mentors: Killian McLoughlin, Rudy Nunez

ABSTRACT: The adverse effects of online sexism have become increasingly challenging to address due to their silencing effect, which is both normalized and subtle. Since behaviors specific to the COVID-19 Pandemic (e.g., wearing face masks) should not differ along gendered lines, any disparity in the genders' outrage would suggest that gendered hostility and punishment are intrinsic to contemporary social life and are not confined to gender-coded behaviors with roots in our past. In other words, hostility is an expression of gender bias rather than a reaction to different behavior by gender. This research examines the differences in the volume and content of toxicity directed towards men and women using a sample of tweets posted to Twitter in English from December 2019 to December 2021 (N= 5,000,000). The LDA topic model results suggested that men received more tweets containing outrage than women. Still, women were targeted with more gendered hostility, including but not limited to gendered slurs and phrases that incite sexual violence. Social media platforms must take preventative measures to limit the spread of toxicity.
KEYWORDS: Behavioral and Social Sciences; Cognitive Psychology; Gender; COVID-19; Online Outrage.

■ Introduction

Social media use has been growing across the globe for more than a decade. Sharing opinions online has two main possible outcomes. The first outcome is the democratizing the public sphere, framing social media as a free, accessible forum that allows for public discourse open to diverse opinions without marginalizing any perspective.[1] The second outcome is the fragmentation of society — small, vocal, public section become the only voices that are heard online.[2] The latter may occur as the result of outrage and moralistic punishment, as humans tend to punish those who deviate from social norms.[3] Research in social psychology suggests that punishment often targets a specific identity, such as the gender identity or race of the recipient.[4] This punishment silences the voices of marginalized groups, thus replicating and even exaggerating existing power structures.[5] A democracy, which relies on openness to diverse perspectives and opinions without the systematic exclusion of certain groups, is threatened, and thus, aggressive online harassment challenges unlimited free speech.[6]

There have been gender differences in compliance with COVID-19 safety recommendations. Men reported significantly lower intentions to wear face coverings in public, self-isolate when sick, avoid public gatherings, and wash hands more frequently.[7] Individuals who decided not to comply with these recommended safety measures were publicly shamed and berated on social media; a phenomenon that the media has dubbed 'pandemic shaming' by influential news outlets such as the New York Times, The Guardian, The Atlantic, and The Washington Post. COVID-19 is a novel context in which to examine possible gendered behaviors. New behaviors required by the pandemic (e.g., wearing face masks) should not have been gendered when they emerged to ensure a fair society with equal treatment. If pandemic-specific behaviors evoked responses that differed along gendered lines, this would suggest that gendered hostility and punishment are intrinsic to contemporary social life and are not confined to gender-coded behaviors with roots in our past. To investigate the differences in online incivility experienced by men and women during the pandemic, I collected social media data from Twitter from December 2019 to December 2021 (N = 5,000,000) and measured the differences in the volume and content of toxicity directed towards men and women.

Past research examining Reddit's r/AmITheAsshole, a thread in which users describe morally ambiguous scenarios in-depth and users vote on whether the Original Poster acted immorally, found that the Original Poster was more likely to be voted the aggressor when they were male.[8] This research suggests that in certain online contexts men receive greater attention for their failure to abide by rules or live up to expectations. However, an exploration of aggressive online communication should not limit itself to the frequency or volume of harassment. The content of such communication must also be assessed. Toxic, outraged, or aggressive comments online may be constructive and designed to correct undesirable behaviors. Still, other comments can incite physical or sexual violence and deploy slurs and hate speech targeting individuals based on their group membership. Aggressive online communication styles may drive marginalized groups, such as oppressed gender identities and racial identities, into silence, measured by a lack of tweets from one demographic about a societal issue.[6] Words such as "bitch" and "slut," among others, target women.[9] If any harassment is gendered at all, whether the frequency is different or not, there is a gender disparity that must be addressed. A possible consequence of this disparity is the silencing of marginalized groups as a response to gendered toxicity.

In this paper, I investigate potential gender disparities in both the amount and content of online toxicity targeting males and females for non-compliance with COVID-19 guidelines. I used a machine learning classifier to measure toxicity in a

DOI: 10.36838/v5i4.6

dataset of tweets posted during the first two years of the pandemic, split by the gender of the target of the tweet. Using a topic model, I also explored whether the content of the tweets expressing toxicity differed across the target gender.

Methods

To test the differences in the volume and content of toxic posts targeting women versus men on Twitter during the pandemic, I collected a novel set of posts made to Twitter from December 2019 to December 2021 (N = 5,000,000). Tweets were collected using a strategy that allowed me to infer the gender of their targets. I also used an existing toxicity classifier (https://perspectiveapi.com) to estimate the probability that each tweet contained toxic language, defined as rude or disrespectful comments likely to lead the recipient to leave a conversation. Using this information, I could compare the counts of toxic tweets targeting men compared to women and test for differences. I also built two topic models, one for each group of tweets; those targeting men and those targeting women. This allowed me to conduct a preliminary investigation into the differences in content or type of language used to target women compared to men.

Dataset:

Twitter is a microblogging and social networking platform where users send out 280-character messages called tweets. Users may read tweets, post tweets, and follow other users to view their tweets. It is currently one of the leading social networking sites worldwide with 237.8 million monetizable daily active users as of 2020.[10] I used the publicly accessible Twitter API to search for tweets that contained references to mask-wearing, hand-washing, or self-isolating alongside gendered pronouns (e.g., she/he). Each tweet had to tag another platform users to be included and posted immediately prior to and during a large portion of the COVID-19 pandemic, December 2019 - December 2021. This methodology allowed me to collect tweets that I could reasonably infer contained content about the unfolding pandemic and which addressed targets identified (at least in part) by their assumed gender. Due to restrictions on the number of tweets that can be collected at any one time from Twitter's APIs, I developed a strategy to sample 5 million tweets equally distributed over the months of interest (i.e., December 2019 - December 2021) and gendered pronouns. My final sample contained exactly 5 million tweets split 50/50 across genders. An exclusive OR logic was used in the queries to Twitter's API so that tweets containing both male and female pronouns were not collected

Operationalizing Toxicity:

I made use of an existing, validated tool for the classification of toxic speech online.[11] Perspective API (https://perspectiveapi.com) was developed by Google to assist content moderators in identifying various types of online text likely to be 'toxic,' where 'toxic' refers to rude and disrespectful language which appears designed to exclude the target from a conversation. Aside from its widespread use in industry and academia, Perspective API was useful for my purposes because it targets language that might be used to exclude users.[11] Given the possibility that online toxicity was used during COVID-19 to exclude certain social groups from conversations about pandemic policy, this operationalization of toxicity can address whether women's voices were sidelined in these debates. Perspective API estimates the probability that it contains severely toxic language for a given piece of online text. I classed as toxic any tweet in my dataset that the Perspective API assigned a probability of being toxic of .51 or higher.

Results and Discussion

Results:

I tested whether there were differences in the volume and content of toxicity targeting men and women on Twitter during the COVID-19 pandemic. Across all the tweets in my dataset, Perspective API classified 18.28% as severely toxic (Figure 1). When split by the gender of the target, a higher proportion of tweets targeting men were classified as toxic compared to those targeting women, 18.5% and 17.68%, respectively (Figure 2). While small, the results of a Fisher's exact test revealed that this difference was significant, $p < .001$, such that men were targeted with a greater proportion of toxicity than women (Odds ratio = 0.95).

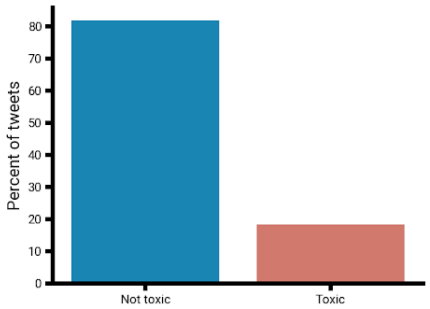


Figure 1: Percentage of tweets that were classified as not toxic and severely toxic by Perspective API.

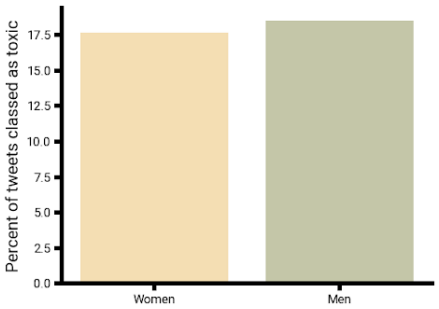
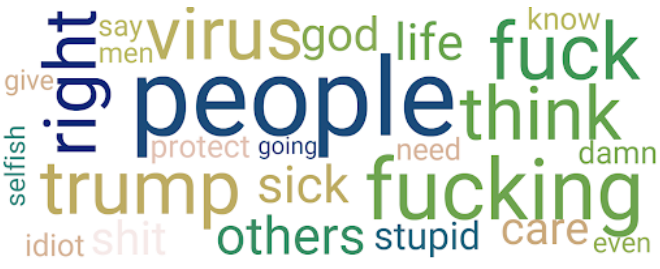


Figure 2: The Proportion of tweets targeting women and men which contained severe toxicity.

I also ran two GSDMM topic models on the portion of the data that was classed as severely toxic.[12] I tested one model on the tweets that targeted men and the other on tweets that targeted women. This allowed me to explore the content-level differences between toxicity targeting men and women. I choose to use GSDMM models because they were explicitly designed to model short texts, such as tweets.[12] Other common topic modeling techniques, such as LDA, perform less well on short texts.[13] There are limited methods to present the results of GSDMM models visually.[14] Below, I present for each model the 25 most frequently occurring words in the topic that contains the most tweets (Figure 3). As seen in the word clouds, the words most associated with men in the

toxic portion of the data invoke general stupidity and terms of abuse (i.e., 'fucking', 'stupid,' and 'idiot'). However, some of the terms associated with women were specifically gendered slur terms (i.e., 'slut', 'bitch,' and 'cunt'). The word 'husband' also appeared in this topic, signaling a woman's relationship to a man. The internet trope of 'Karen' was also referenced. The specificity of the toxicity in the second model suggests content-level differences in tweets targeting men and women. Women appeared more likely to be targeted with gendered slurs and indexed to their husbands or to an internet meme of an entitled woman.

A)



B)

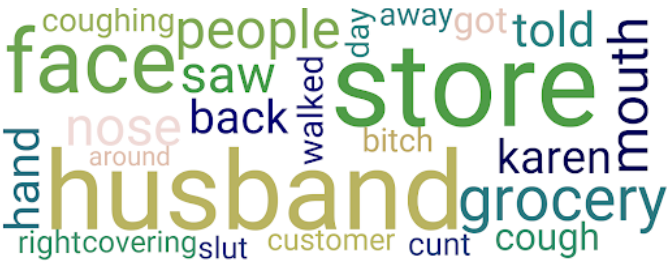


Figure 3: The 25 most frequently used words in the topic that contained the most tweets according to GSDMM models run on tweets that contained severely toxic language and that targeted (A) men and (B) women.

Overall, these results suggest that men received more tweets containing toxic language than women, but women were targeted with more gendered slurs.

Discussion

In this research, I hypothesized that 1) women would receive a higher volume of tweets containing toxicity, and 2) women would receive harsher and more offensive toxicity. The results surprisingly only supported the second hypothesis. However, the finding that women receive more gendered slurs than men while receiving fewer toxic tweets overall indicates that the gender disparity is stronger and more entrenched than the hypothesis may have predicted, urgently suggesting the necessity for political action.

While this research offers new insights and supports existing research, certain factors that limit my conclusions. For example, determining the gender identity of the users was limited to making inferences based on the pronouns she/her and he/him. Because they/them is both singular and plural, these data are too ambiguous to analyze and make conclusions about toxicity toward people who do not use she/her or he/him pronouns. Twitter does not provide exact demographic information to researchers. Additionally, this study only looked at toxicity pertaining to following safety guidelines and protocols for COVID-19, not all toxicity in the vast world of social media. This limits the generalizability of my findings to other domains.

Perhaps social media hate speech filters should add more warning signs, scanning for hate speech or language that incites violence before a tweet or allow users to filter those sentiments out of their feeds. While there exists a report feature for every post, the hateful speech is already online, creating a negative effect, making the change a reaction instead of a preventative measure.

Further research should focus on understanding why women receive more gendered and offensive posts online. Researchers should continue to suggest how to dismantle entrenched societal sexism and how society may work toward making outrage constructive instead of marginalizing.

Thus, although this research does not definitively provide a solution to women experiencing more gendered toxicity it exposes a problem to be addressed, one that likely affects all social media users.

Conclusion

While men receive a greater number of tweets containing outrage, women receive more gendered hostility in the content of the tweets, such as slurs that are offensive to women or phrases that incite sexual violence. This research demonstrates that sexism and punishment in society are still imposed by the community and are not yet obsolete, requiring action to be taken. To curb the negative effects, social media platforms should harness their influence to ensure that no voices are marginalized so that all may contribute to democratic public discourse.

Acknowledgments

I am deeply grateful to Killian McLoughlin, a Princeton Ph.D. student, for guiding me and offering resources for me to analyze my data. I am also appreciative of my teacher at Riverdale Country School, Rudy Nunez, for supporting and igniting my passion for pursuing data science to find out the truth about the world.

References

1. Ash, T. Free speech. Ten principles for a connected world. London, England: Atlantic Books 2016.
2. Bright, J. Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, 23, (1), January 2018, Pages 17–33. https://doi.org/10.1093/jcmc/zmx002.
3. Feng, C.; Luo, Y.; Krueger, F. Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. *National Library of Medicine* 2014, 36. (2): 591-602. https://doi.org/10.1002/hbm.22649.
4. Clark, M.; Lemay, E.; Reis, H. Other People as Situations: Relational Context Shapes Psychological Phenomena. *The Oxford Handbook of Psychological Situations* 2020, online edn. https://doi.org/10.1093/oxfordhb/9780190263348.013.5.
5. Nadim, M.; Fladmoe, A. Silencing Women? Gender and Online Harassment. *Social Science Computer Review* 2021, 39, (2).https://doi.org/10.1177/0894439319865518.
6. Nadim, M.; Fladmoe, A. Silenced by hate? Hate speech as a social boundary to free speech. Boundary Struggles. *Contestations of Free Speech in the Public Sphere* 2017, pp. 45-76.
7. Everett, J.; Colombatto, C.; Chituc, J,; Brady, W,; Crockett, M. The Effectiveness of Moral Messages on Public Health Behavioral Int-

DOI: 10.36838/v5i4.6

entions During the COVID-19 Pandemic. *Psychology* 2020, 3 (1), 4. psyarxiv.com/9yqs8.
8. Botzer, N,; Gu. S.; Weninger, T. Analysis of Moral Judgment on Reddit, *IEEE Transactions on Computational Social Systems* 2021. https://doi.org/10.1109/TCSS.2022.3160677.
9. Felmlee, D.; Rodis, P.; Zhang, A. Sexist Slurs: Reinforcing Feminine Stereotypes Online. *Sex Roles* 2020, 83, (1), 16–28. https://doi.org/10.1007/s11199-019-01095-z.
10. Statista. https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/.
11. Rieder, B.; Skop, Y. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 2021, 8(2). https://doi.org/10.1177/20539517211046181.
12. Yin, J.; Wang, J. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* 2014.
13. Albalawi, R.; Yeap, T. G.; Benyouce, . Using topic modeling methods for short text data: A comparative analysis. *Frontiers in Artificial Intelligence* 2020, 3. https://doi.org/10.3389/ frai.2020. 00042.
14. Pelgrim, R. Short-test topic modelling: LDA vs GSDMM. TowTards Data Science 2021. https://towardsdatascience.com/short-text-topic-modelling-lda-vs-gsdmm-20f1db742e14.

Author

Christine Lee is an eager data scientist to find impactful solutions to societal problems, specifically those related to the oppression of marginalized groups. Outside of research, Christine is interested in playwriting, a passion that complements data science because both require an understanding of influences that skew perspectives.