Public's demand for moderation to online toxic speech is limited and influenced by social factors*

Yichen Ji Xiaoxu Liu

February 13, 2024

In this report,we analysis the public's ideal moderation for 5 levels toxic online speech.It was found that most people tend to not give harsh moderations, unless the speech touches upon personal threat. And people with different social factors have difference preferences on moderations.The findings will help me build a ideal standard platform for the public.

Table of contents

*Code and data are available at: https://github.com/Selinayichenji/Toxic_speech_replication.git.

1 Introduction

*This reproduction was performed after a replication on the Social Science Reproduction platform:**link here***

The research has shown that roughly 4 in 10 Americans have experienced online harassment, including name-calling, physical threats, and sexual abuse (Vogels 2021). With the current widespread use of social media, the debation of regulating toxic content becomes even more pertinent. Cultivating civility within democratic discourse is strongly necessary, such as emphasizing respect and social order when communicating online. However, the emergence of uncivil, intolerant content on social platforms raises concerns about its potential harm to public discourse and democracy. Our focus is on the crucial dilemma: should measures be implemented to moderate toxic content and uphold civility, or should we suggest allowing such speech on social media to remain unconstrained? The pursuit of addressing these challenges becomes significant in shaping the future of online democratic engagement.

Although hate, harassment, and extremism significantly impact the country and online community negatively as toxic comments have saturated lots of common social media in the U.S., the application of strong and effective regulation over social media faces challenges due to multiple factors. Various factors, including technology companies, government, and NGOs, oppose the potential heavy regulation, which operates within a distinct legal framework in the U.S. Besides, Users, the ultimate recipients of online toxicity, are important in reporting objectionable content through flagging mechanisms. Therefore, we attempt to find the ideal platform standards from users' views to against toxic speech.

We aim to apply the initial analysis from the original paper "Toxic Speech and Limited Demand for Content Moderation on Social Media", which is from the American Political Science Review. The paper attempts to figure out the consistency of how users reply to toxic speech when using social media to find an appropriate solution to improve the harmony of online platforms, and it includes two pieces of research: 1. targeting social groups and 2. targeting partisans.

In our reproduction, we use the original methods and the same dataset and shift the concentration from the level of toxic speech to the types of users while extending the research with more specific prevalent aspects of the groups of people's responses. Instead of talking about how people respond to toxic speech toward different labels of victims (LGBTQ, billionaire, and Christian) in study 1 of the original paper, we focused on how people with different genders, education levels, and races react to toxic speech toward LGBTQ. The estimand of our paper is

differences of tendency of giving content moderation among people with different gender, race and education level when they face the toxic speech's target is LGBTQ.Beyond the changes, we hold all other perspectives to be the same as the original paper.

We obtain the result of the reproduction work and find that white people, people have post-graduate degree and man tend to give slighter moderation than other groups. Woman and black people tend to give serious moderation than others. The research result provides us with the information that: - In the traditional sense, socially advantaged groups tend to propose more lenient content moderation measures when the LGBTQ community is attacked, whereas disadvantaged groups are inclined to penalize speech that attacks the LGBTQ community. - People usually want a more loose-controlled online environment and choose no heavy moderation toward heavy speech.

These help us understand the user preference in social media, and stimulate the appropriate regulations of the speeches on online platforms. Generally, the reproduction talks about the summary of what we do based on the original paper and what we obtain, the data sources, detailed pictures and analysis through coding, and the discussion that concludes our results and lessons while discussing (potential) drawbacks and anticipated regulations/behaviors in the future.

The paper introduces the basic information contains data source, methodology, variables and measurements in Section 2. The visualization and analysis of the tendency differences in moderation by people across three social factors are presented in Section 3. And in Section 4, we discuss what we have learned, our understanding of the world, the limitations, and the next steps of our research. Section A is for appendix and Section A is listed all references in this paper.

2 Data

2.1 Source

Our replication paper is based on the original paper in American Political Science Review "Toxic Speech and Limited Demand for Content Moderation on Social Media". Our paper is consistent with the original goal that attempts to find how people respond to toxic speech with different targets to decide the moderation of social media for a respectful and harmonious environment. The dataset and replication package is open on the Harvard Dataverse website. For the data in LGBTQ target, the only esstential one is lgbtq.RData in the replication package.

The data came from the survey results collected by the authors of original paper. So there is no other similar datasets for same research.

2.2 Methodology

Our paper applies using the statistical programming language R (R Core Team 2022). Besides the programming tool, we also employ the following packages: readr (Wickham, Hester, and Bryan 2023), broom (Robinson, Hayes, and Couch 2023), ggplot2 (Wickham 2016), dplyr (Wickham et al. 2023), tidyverse (Wickham et al. 2019), MetBrewer (Mills 2022), and knitr (Xie 2023).

2.3 Variables

We only introduced the variables used in our own analysis. For the full variables in survey from the original paper, please check the Appendix.

- Treatment: non-group-related control, control, uncivil, intolerant and threatening.
  - Non-group-related control means no target and no toxic language.
  - Control means anti-target but without the 3 kinds of toxic languages: uncivil, intolerant and threatening.
  - Uncivil was defined as "including anything from an unnecessarily disrespectful tone and lack of respect to rudeness and inconsiderate language."(PRADEL F, n.d.).
  - The intolerance differs from incivility, "it aims to derogate, silence, or undermine particular groups due to their protected characteristics, attack their rights, and incite violence and harm."(PRADEL F, n.d.).
  - The threatening is a toxic behavior explicitly announces the intention of physical harm.(PRADEL F, n.d.).
- Handle: (1) Leave it, do nothing; (2) Place a warning label on the post; (3) Reduce how many people can see the post; (4) Permanently remove the post; (5) Suspend the person's account.
- Gender: Male, Female and Others.
- Education: High school graduate; College; Postgraduate.
- Race: White, Black, Hispanic and Others.
- Percentage: It means the ratio of 5 handles given by people with the same social factor when they are facing the same treatment.

2.4 Measurement

The data collected by using survey. In the study I (target social group), targets include LGBTQ, Billionaire and Christians. The study measures participants' preferences for handling toxic speech on social media by asking them to react to posts on social media platforms. Specifically, the research measures participants' preferences through the following question:

- "In your view, how should social media companies (such as Facebook and Twitter) handle the post above?" Participants can choose from the following options:
  - "Leave it, do nothing"
  - "Place a warning label on the post"
  - "Reduce how many people can see the post"
  - "Permanently remove the post"
  - "Suspend the person's account"

These options allow participants to express their preferences for the actions social media platforms should take regarding toxic speech. By observing participants' choices among these options, researchers can understand participants' varying preferences for content moderation on social media platforms under different experimental conditions.

The experiments took place in July 2022 (LGBTQ and Billionaires) and October 2022 (Christians). Each study recruited between 1,300 and 2,000 U.S. adults from the participant pool of the crowdsourcing platform Prolific. Participants were recruited according to specific procedures outlined in the APSR Dataverse (Pradel et al. 2024). Exclusion criteria, as pre-registered, included participants who failed the attention check, opted for exclusion from the study, or completed the survey in less than 50 seconds. Ethics approval was obtained from the Central University Research Ethics Committee of the University of Oxford, and the study design was pre-registered prior to data collection.(PRADEL F, n.d.)

3 Results

We selected 3 social factors (gender, education level and race) to investigate people's moderation preference in LGBTQ Target toxic speech.

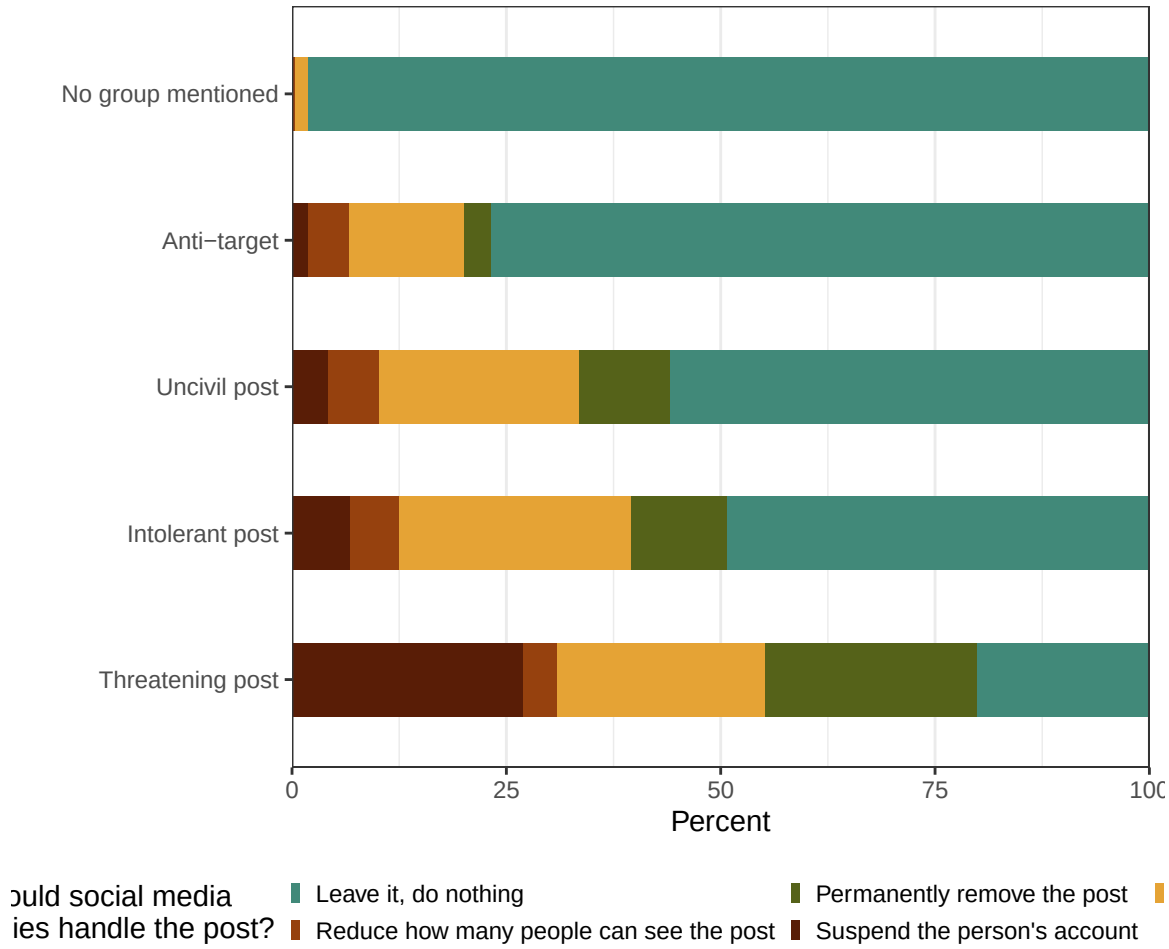

Figure 1

Figure 1 illustrates the general percentages of 5 different handles in 5 levels toxic speech targets LGBTQ . As shown on the Figure 1, the serious content moderation takes more percentage as the toxic level increases. The interviewees are especially sensitive to threatening post to LGBTQ people. For no group mentioned post, most people choose to leave it, a few people put a warning. But in Threatening post, the percentages of permanently remove the post and suspend the account increase obviously, and that of other 3 decrease compared to Intolerant post, which means people choose to let those posts disappear instead other indirect moderation.

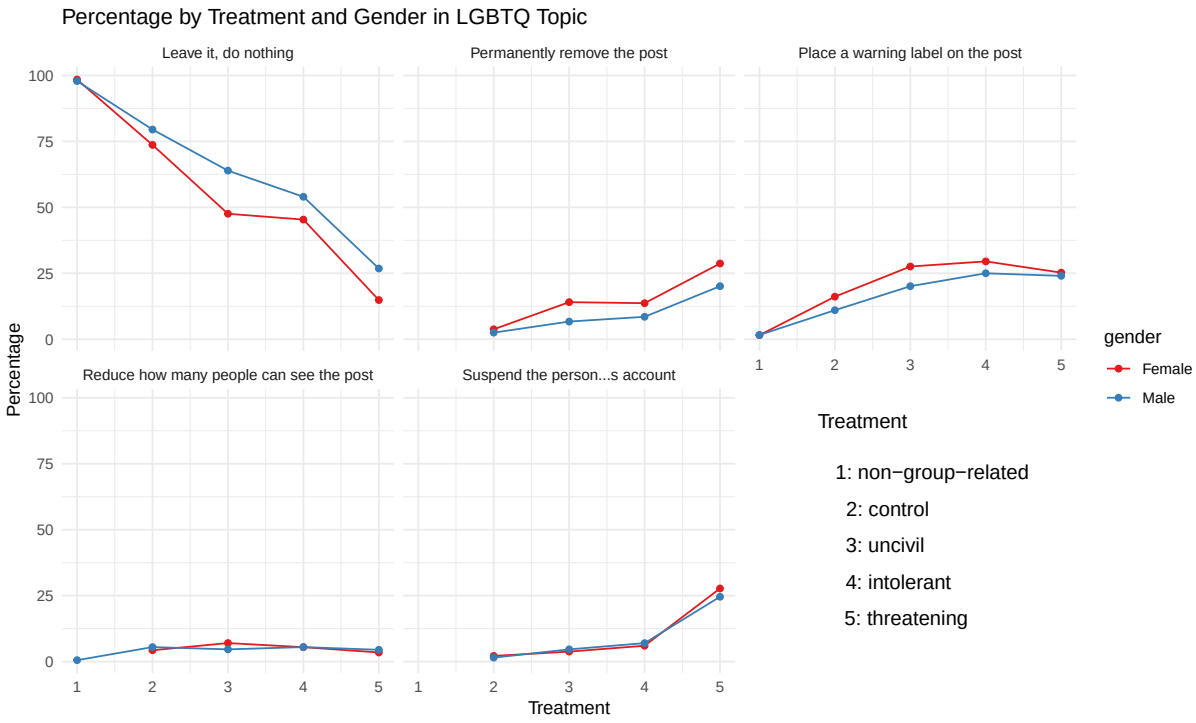

Figure 2

In Figure 2, Figure 3 and Figure 4, we visualize the difference of tendency of different groups' moderation. The plots arranges in order of the seriousness of moderation.

From Figure 2, female tends to give more harsh moderation than male. The percentage of slightest moderation (leave it, do nothing) in male group is about 12.5% higher than that of woman in all levels toxic speech. Oppositely, the percentages of permanently remove the post and place a warning label in female group are obviously higher than male about 7%. To reduce post' view and suspend account, there is no obvious difference between 2 groups.

As shown on Figure 3, white people tend to give slighter moderation than other races. Black people tend to give the harshest moderation among all races. In Figure 4, people with postgraduate degree tend to give slighter moderation on toxic speech. No obvious difference between College and High school graduates groups.

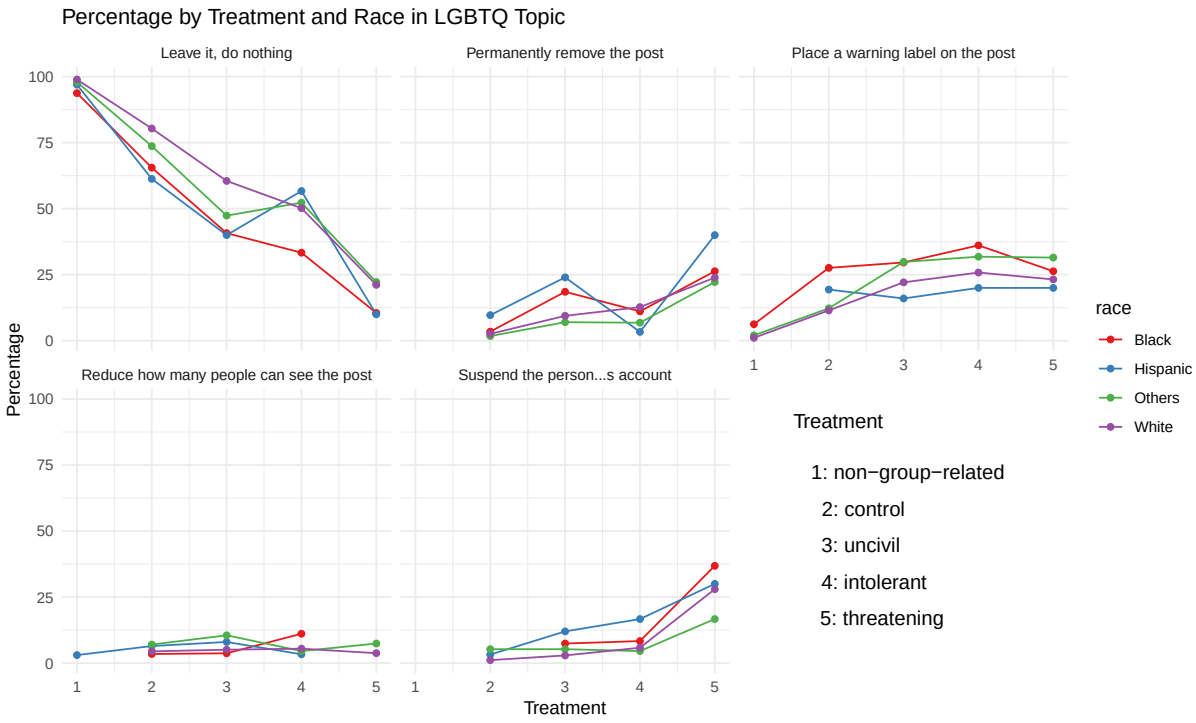

Figure 3

4 Discussion

4.1 What we did

We replicated 3 figures contains 2 panels "Target LGBTQ" and "Target: Billionaires" in Figure3 "Preferences for Content Moderation by Treatment and by Experiment in Study I" and 1 element, Figure 6 "Preferences for Specific Types of Content Moderation by Treatment in Study II" from the paper "Toxic Speech and Limited Demand for Content Moderation on Social Media". Only the "Target LGBTQ" is shown in the paper as Figure 1. The other figures' replication are in the scripts folder in Github, please check the link in the first page.

Different from the original paper, we focused on the 3 social factors groups' difference of preference on content moderation by treatment. By using the percentages' differences, we avoids the issue of

4.2 What we acquire

In the paper, the result represents a trend that users prefer to do nothing to the toxic speech even if they may be affected negatively. The first study indicates that people are more likely

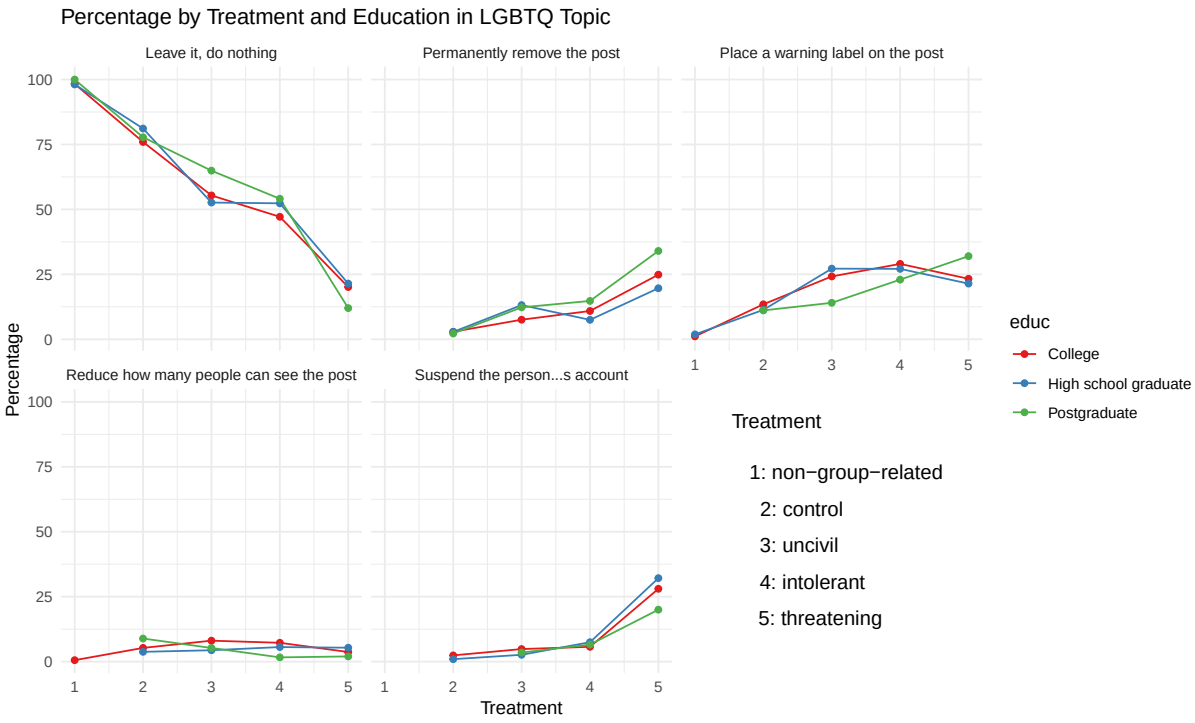

Figure 4

to react to a toxic speech by reducing how many people can see the post or suspending the account if the target is LGBTQ and there is not much difference if the targets are billionaires or Christians. The second study implies that Democrats may tend to protect LGBTQ more while Republicans care about Christians more. In contrast with these, our reproduction acquires more detailed conclusions from three slightly different perspectives. From gender sight, women show a stronger tendency to protect LGBTQ than men do. From the education level perspective, those who are high school graduates tend to exhibit more protective behaviors when dealing with higher levels of toxic speech while those who are postgraduate act more sensitively to the uncivil levels. From the race perspective, Blacks are usually more sensitive to the LGBTQ target overall.

4.3 What we learn: Point 1: Diverse Consequences of Different Toxic Speech Types

The study challenges past research by revealing various types of toxic speech, including incivility, intolerance, and violent threats. Unlike previous work that often focused solely on labeling manifestations of incivility, this research connects toxic speech types to their consequences. It introduces a nuanced perspective, suggesting that users perceive these speech types as distinct constructs. While intolerance and incivility prompt similar content moderation responses, the

study highlights the empirical insight that users view them differently, considering incivility a matter of tone and intolerance a matter of substance, such as discrimination.

4.4 Point 2: Limited Support for Content Moderation

A significant finding is the overall low support for content moderation of uncivil and intolerant content. The majority of respondents express the view that such content should remain online, with censorious forms of moderation, like banning users or removing content, being among the least favored options. Even when presented with extreme cases of toxic speech, such as attacks on the LGBTQ community, a large portion of respondents do not advocate for content moderation. This raises concerns about the broader implications for public discourse, as a substantial portion of users seems reluctant to endorse content moderation, even in the face of highly objectionable speech.

4.5 Point 3: Partisan Consistency and New Research Avenues

Contrary to expectations in an era of affective polarization, the study finds limited evidence that users view moderation of toxic speech through partisan lenses. Democrats, in general, show a greater tendency to demand moderation, but the identity of the victim (Republican or Democrat) does not significantly influence partisans' views. This finding opens up a new research puzzle, suggesting that Americans' strong belief in the value of freedom of speech might be a driving factor. The study calls for further exploration into whether this trend persists in other countries with different legal frameworks. The results emphasize the need to understand content moderation preferences beyond partisan lines and suggest that Americans, despite political polarization, exhibit consistency in their views on this matter.

4.6 Weaknesses

The primary limitation of our research stems from the lack of modeling. The absence of models means that we do not account for potential interactions between different social factors and how they may collectively influence moderation behaviors. This gap is significant because different social groups may have varying degrees of influence on content moderation decisions, and without proper weighting in our analysis, the results might not fully reflect the complexities of these interactions.

Moreover, another limitation is related to the representativeness of our sample. The percentages of different social factor groups in our survey may not accurately mirror the actual composition of society. This discrepancy can lead to skewed results, as the moderation tendencies we observe may be over or under-represented due to sampling bias. For example, if a social factor group that is more likely to advocate for stricter moderation is under-represented in the survey, our findings may underestimate the desire for such moderation across the population.

The survey results involve the interpretation of social tendencies toward content moderation, which are inherently nuanced and subject to the cultural and social dynamics within each group. However, our current approach does not allow us to explore these dynamics in depth, which could provide valuable insights into the reasons behind the observed moderation preferences.

4.7 Next steps

To address these limitations, our next steps involve the introduction of statistical models, such as regression analyses, which can help us understand the weight and impact of each social factor on moderation decisions. Regression models would allow us to control for various covariates and examine the independent effect of each variable. By doing so, we can also investigate interaction effects, providing a more nuanced understanding of how multiple social factors may interplay to influence content moderation behaviors.

Additionally, efforts should be made to ensure that our sample is more reflective of society's actual demographic composition. This could involve stratified sampling or weighting survey responses to match the demographic makeup of the population. Implementing these methods will likely result in more generalizable and accurate findings that can better inform content moderation policies and practices.

A Appendix

References

Mills, Blake Robert. 2022. *MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art*. https://CRAN.R-project.org/package=MetBrewer.

PRADEL F, KOSMIDIS S, ZILINSKY J. n.d. "Toxic Speech and Limited Demand for Content Moderation on Social Media." CAmerican Political Science Review. https://doi.org/10.1017/S000305542300134X.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Vogels, Emily A. 2021. *The State of Online Harassment*. Pew Rsearch Center. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2021/01/PI_2021.01.13_Online-Harassment_FINAL-1.pdf.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.