Public's Ideal Moderation to Online Toxic Speech*

Yichen Ji Xiaoxu Liu

February 12, 2024

In this report,we analysis the public's ideal moderation for 5 levels toxic online speech.It was found that most people tend not to 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

1 Introduction **2**

2 Data **3**
2.1 Source . . . 3
2.2 Methodology . . . 3
2.3 Variables . . . 4
2.4 Measurement . . . 4
2.5 Structure of the paper'remainder parts . . . 4

3 Results **5**

4 Discussion **8**
4.1 What we acquire . . . 8
4.2 What we learn: Diverse Consequences of Different Toxic Speech Types . . . 8
4.3 Point 2: Limited Support for Content Moderation . . . 8
4.4 Point 3: Partisan Consistency and New Research Avenues . . . 9
4.5 Weaknesses and next steps . . . 9

Appendix **10**

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

References **11**

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 (Pew Research Center 2021a). 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 expand the study range to how people with different genders, education levels, and races react to toxic speech toward LGBTQ, billionaires, and so on. Beyond the changes, we hold all other perspectives to be the same as the original paper. For example, we set standards of toxic speech with different levels: incivility (disrespectful tone, lack of respect, rudeness, and inconsiderate language), intolerance (derogating, silencing, or undermining particular groups with specified labels), and violent threats(the tendency of physical harm).

We obtain the result of the reproduction work and find that women (gender), high school graduates (education), and Blacks / Hispanics (races) imply a tendency to be more sensitive to high-level toxic speech on social media; on the other side, men, high school graduates, and Blacks tend to be more sensitive to the billionaires target overall. The research result provides us with the information that 1). Different groups of people hold different attitudes and behaviors toward topic speech with different targets and 2). 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. (can be improved)

ESTIMATE

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.

This paper applies the same sources as the original paper does, which consists of Papacharissi (2004, in particular pages 261–7), Herbst (2010), and Boatright et al. (2019); national laws (Busch 2022; European Parliament 2022); Brandeis notes in Whitney v. California 1927, "Partisan Conflict over Content Moderation Is More Than Disagreement about Facts." (Appel, Pan, and Roberts Reference Appel, Pan and Roberts 2023; Kozyreva et al. Reference Kozyreva, Herzog, Lewandowsky, Hertwig, Lorenz-Spreen, Leiser and Reifler 2023), BBC. 2012. "Reddit Will Not Ban 'Distasteful' Content, Chief Executive Says", Bejan, Teresa M. 2017. Mere Civility. Cambridge, and so on.

Similar dataset

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 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. In original paper, uncivil is defined as "including anything from an unnecessarily disrespectful tone and lack of respect to rudeness and inconsiderate language."(PRADEL F, n.d.). And 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.

2.5 Structure of the paper'remainder parts

3 Results

Our results are summarized in **?@tbl-modelresults**.

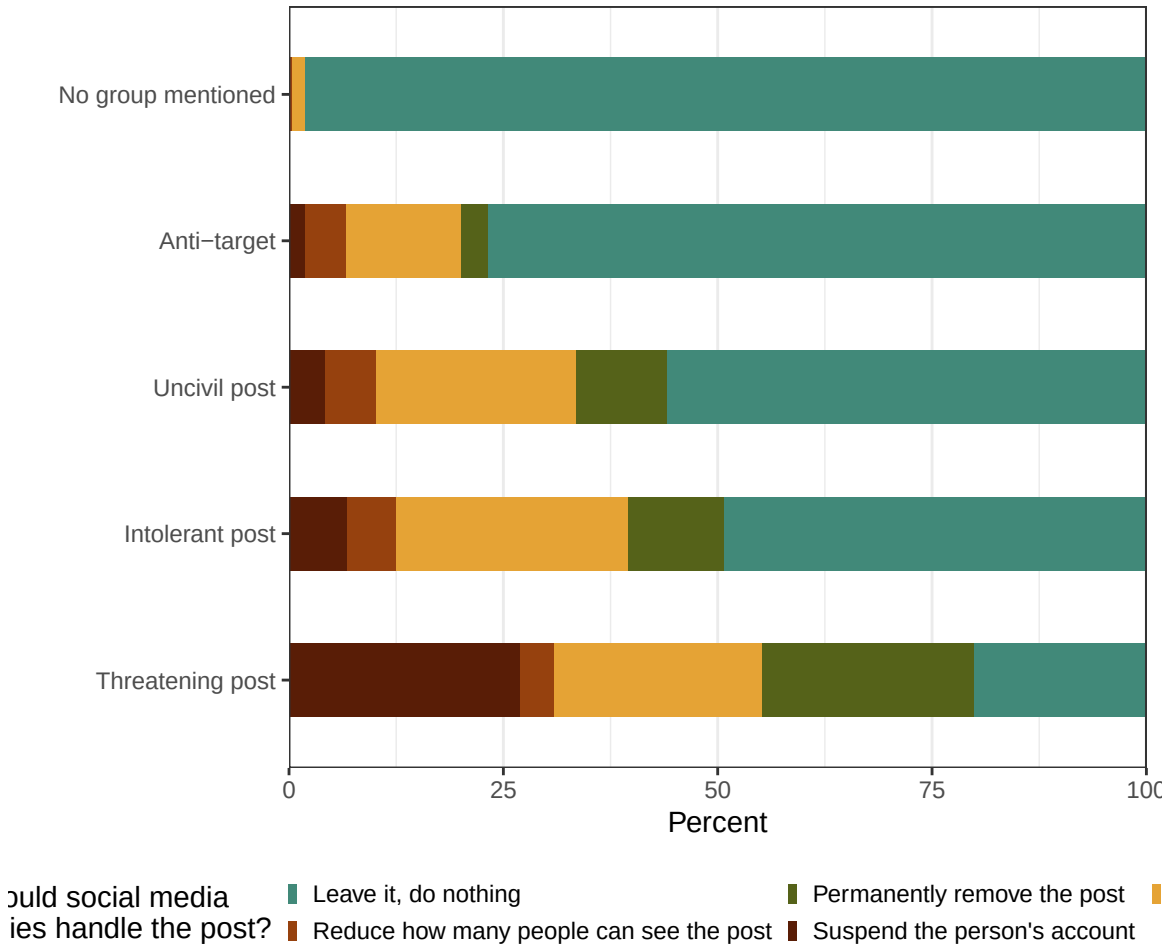
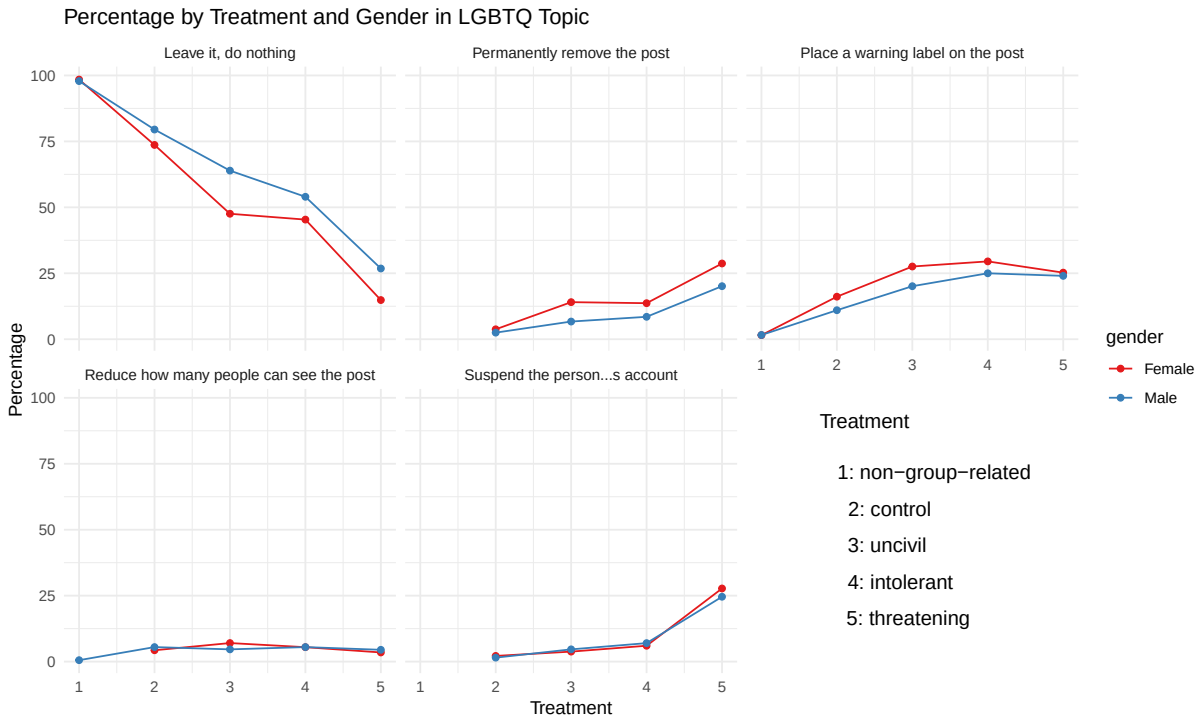
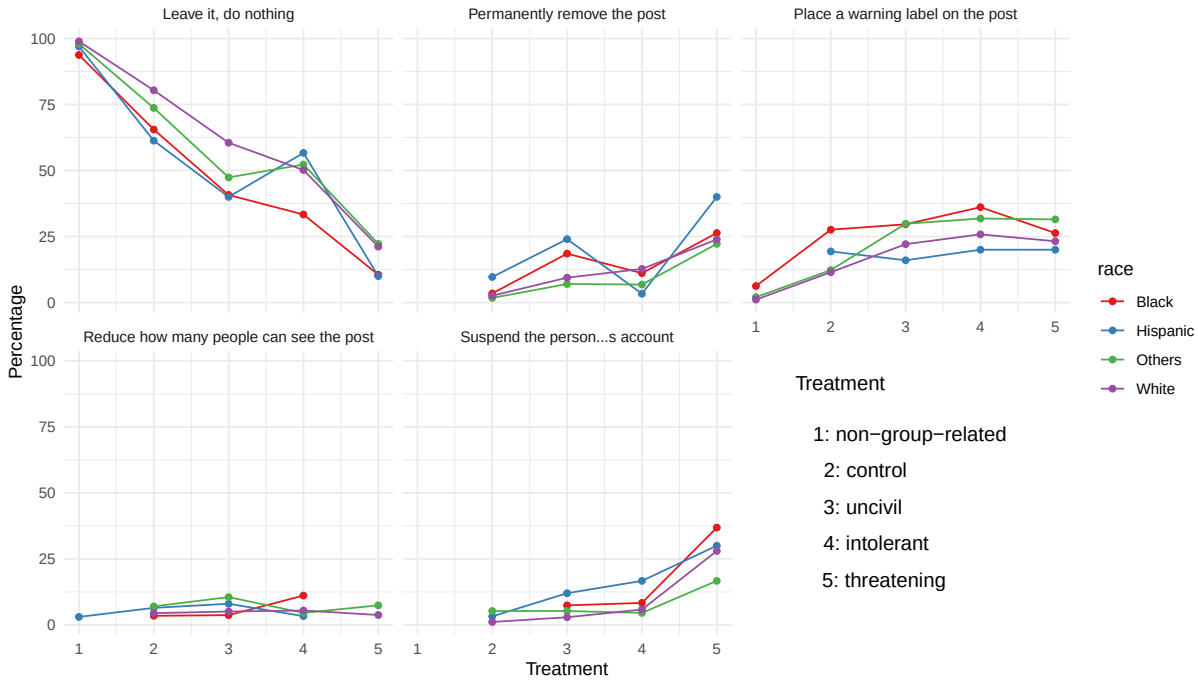
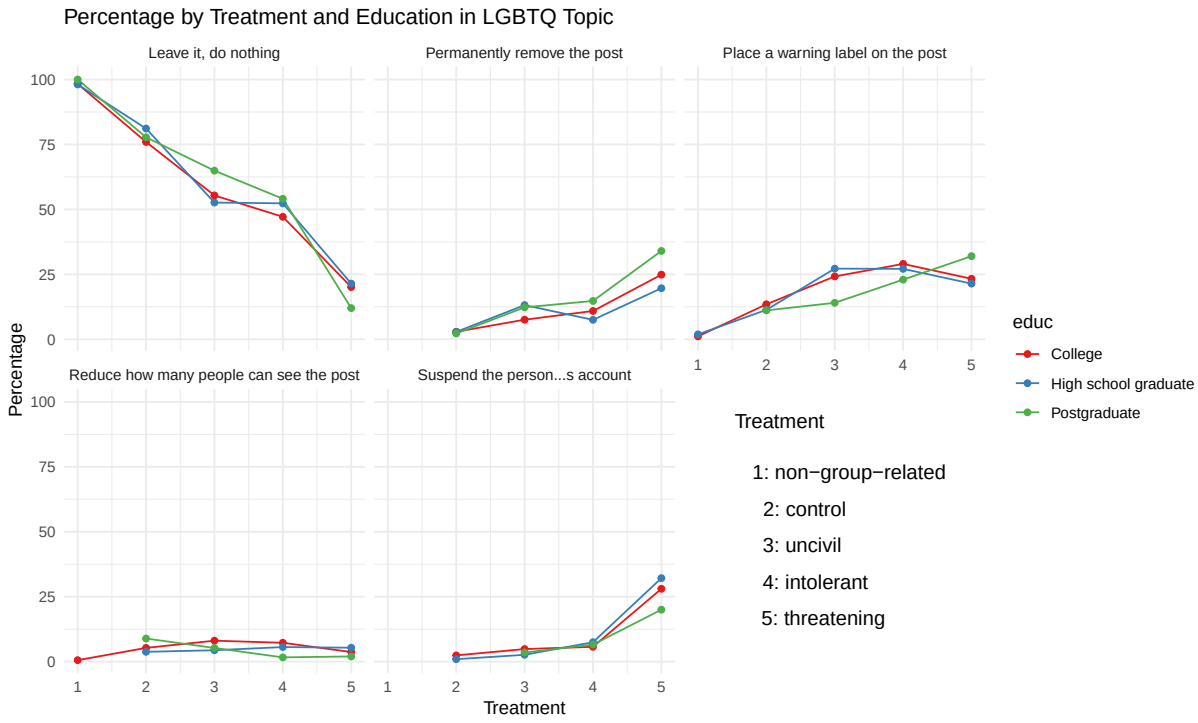


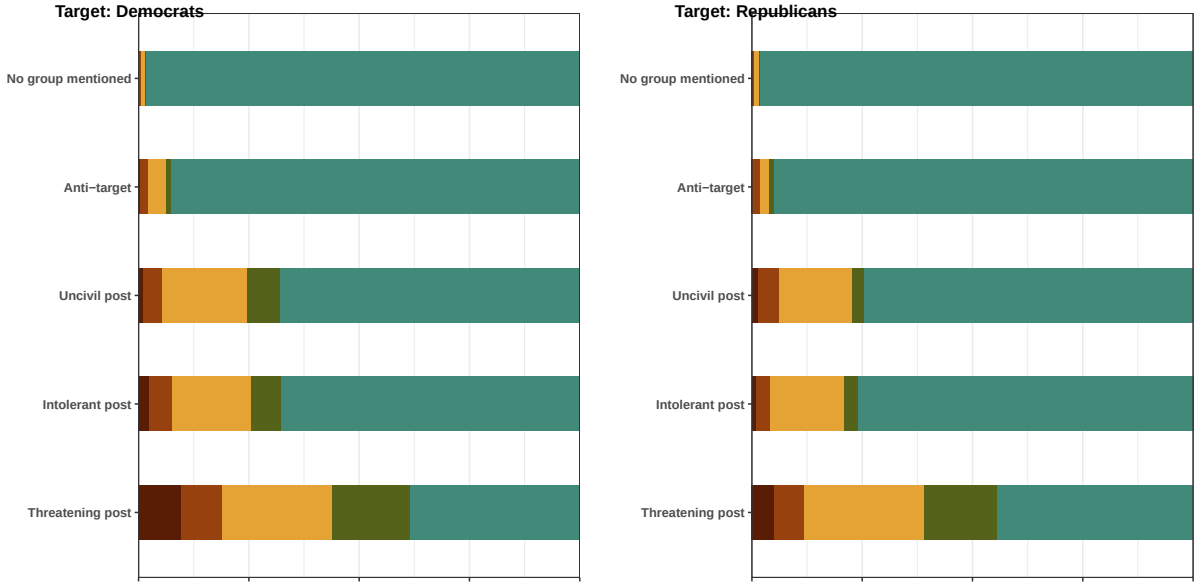
Figure 1

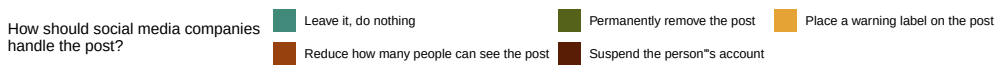
Figure 1


Percentage by Treatment and Gender in LGBTQ Topic
Leave it, do nothing
Permanently remove the post
Place a warning label on the post
Reduce how many people can see the post
Suspend the person...s account
Percentage
Treatment
gender
Female
Male
Treatment
1: non−group−related
2: control
3: uncivil
4: intolerant
5: threatening


Percentage by Treatment and Race in LGBTQ Topic
Leave it, do nothing
Permanently remove the post
Place a warning label on the post
Reduce how many people can see the post
Suspend the person...s account
Percentage
Treatment
race
Black
Hispanic
Others
White
Treatment
1: non−group−related
2: control
3: uncivil
4: intolerant
5: threatening


Percentage by Treatment and Education in LGBTQ Topic
Leave it, do nothing
Permanently remove the post
Place a warning label on the post
Reduce how many people can see the post
Suspend the person...s account
Percentage
Treatment
educ
College
High school graduate
Postgraduate
1: non−group−related
2: control
3: uncivil
4: intolerant
5: threatening


Target: Democrats
Target: Republicans
No group mentioned
Anti−target
Uncivil post
Intolerant post
Threatening post


How should social media companies handle the post?
Leave it, do nothing
Permanently remove the post
Place a warning label on the post
Reduce how many people can see the post
Suspend the person"s account

4 Discussion

4.1 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 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, men tend to agree with higher punishment than women do toward billionaires from gender sight, while 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 and Hispanics are usually more sensitive to the LGBTQ and billionaires target overall.

4.2 What we learn: 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.

4.3 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.

4.4 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.

4.5 Weaknesses and next steps

Weaknesses and next steps should also be included.

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.

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/.