Countering Sexist Hate Speech on YouTube: The Role of Popularity and Gender

Jaeung Sim
KAIST College of Business
Seoul 02455, Republic of Korea
jaeung@kaist.ac.kr

Jae Yeon Kim
University of California, Berkeley
Berkeley, CA 94704 USA
jaeyeonkim@berkeley.edu

Daegon Cho
KAIST College of Business
Seoul 02455, Republic of Korea
daegon.cho@kaist.ac.kr

Extended Abstract

While social media have empowered individuals' voices, they have also amplified adverse influences of such voices on other people. For instance, a hate-speech comment in social media can spread globally and hurt numerous people who did not relate to the hate-speech speaker. Furthermore, it may affect the audience's attitudes and behaviors, causing a long-term damage to the society. To discourage online social speech, recent studies explored a counter-speech strategy, directly blaming hate-speech speakers for verbal violence. Although these studies showed how counter speech silenced such speakers, the audience's responses—the potential loudspeaker of hate speech—have been neglected. In this study, we examined the audience responses to counter speech in the context of sexist hate speech on YouTube, focusing on the moderating role of the popularity of the countering comment and the messenger's gender. Our findings from an online experiment with a sample of 1,250 adult citizens in South Korea showed that counter speech encouraged the audience's intention to report hate-speech comment to YouTube significantly, only when the counter-speech reply received only a few upvotes from other users and was written by a woman. Notably, counter speech did not affect the audience's attitude toward the hate speech, the counter speech, and their messengers. We also found a significant drop in reporting intention among young adults, when counter speech with many upvotes was provided. Lastly, we found that counter speech with small number of upvotes positively shifted the audience's attitude toward both internal and external regulation policies. These results provide insights on how supports for counter speech can backfire in persuading the audience and how social media platforms can effectively facilitate self-correction in user-generated content.

Keywords

Online hate speech, sexist hate speech, counter speech, social media

Countering Sexist Hate Speech on YouTube: The Role of Popularity and Gender

Social media have enabled individuals, who were underrepresented in the mass media era, to upload their own content on public forums or their own channels for free (Susarla et al. 2012). They have created various social values such as fostering civic engagement, empowering social movements, and disciplining corrupt companies (Enikolopov et al. 2018, Freelon et al. 2018, Oh et al. 2015, Warren et al. 2014). The democratic nature of social media, however, has also increased access to inappropriate content. Large-scale social networks and sharing functions such as hashtags have accelerated the diffusion of hate-speech content and unverified rumors (Fox et al. 2015, Shin et al. 2017). Complete or partial anonymity have inhibited constructive discussions through encouraging abusive language (Cho and Kwon 2015). Furthermore, such extreme content may affect the audience's attitudes and behaviors, causing a long-term damage to the society. For instance, swearing in online comments generally increase user attention to as well as other users' approvals to the user-generated content (Kwon and Cho 2015).

Facing this challenge, social media platforms have attempted to reduce abusive language in various ways. For instance, Facebook (2020) has officially banned hate speech, violence and graphic content, false news, and many other types of malicious content. Similarly, YouTube (2020) has not allowed violent or dangerous content such as hate speech, harassment and cyberbullying, and violent or graphic content. In addition to the spontaneous actions of these platforms, some administrations introduced to policies mandating online platforms to make rapid responses to such violation. France, for example, passed the law forcing online platforms to remove within 24 hours hate-speech content (TechCrunch 2020b).

Because countless posts are being uploaded in real time, and the current machine learning algorithms cannot completely detect all malicious content, such platforms in part rely on the reporting systems wherein users declare inappropriate content to the service operators. When sufficient number of reports on certain content arrive, the platforms review the content and remove it upon the appropriateness of the reports. Although user reporting is crucial to maintain the quality of content, few studies examined the motivation of and the way to promote reporting behaviors in social media. Based on focus-group interviews, Johnson (2018) suggested that social media users were not strongly motivated to censor extreme speech and expressed apathy and cynicism toward both their own and social media companies' ability to combat such speech. In other words, many users do not take actions to reduce inappropriate content even though they can actually contribute to the entire platforms.

In this study, we aimed to examine whether counter speech, a widely-adopted strategy that directly sanction hateful or harmful speech (Mathew et al. 2019), can motivate the audience to report sexist hate speech. Prior studies have shown that counter speech effectively discourage harassers in some conditions (e.g., Munger 2017, Siegel and Badaan 2020), but they provide few insights on how the audience changes their attitudes toward hate speech and responses to the counter speech. Furthermore, most of these studies have neglected how to address sexist hate speech, while numerous studies have revealed the prevalence of sexism in social media (e.g., Döring and Mohseni 2020, Nakandala et al. 2017, Wu 2019).

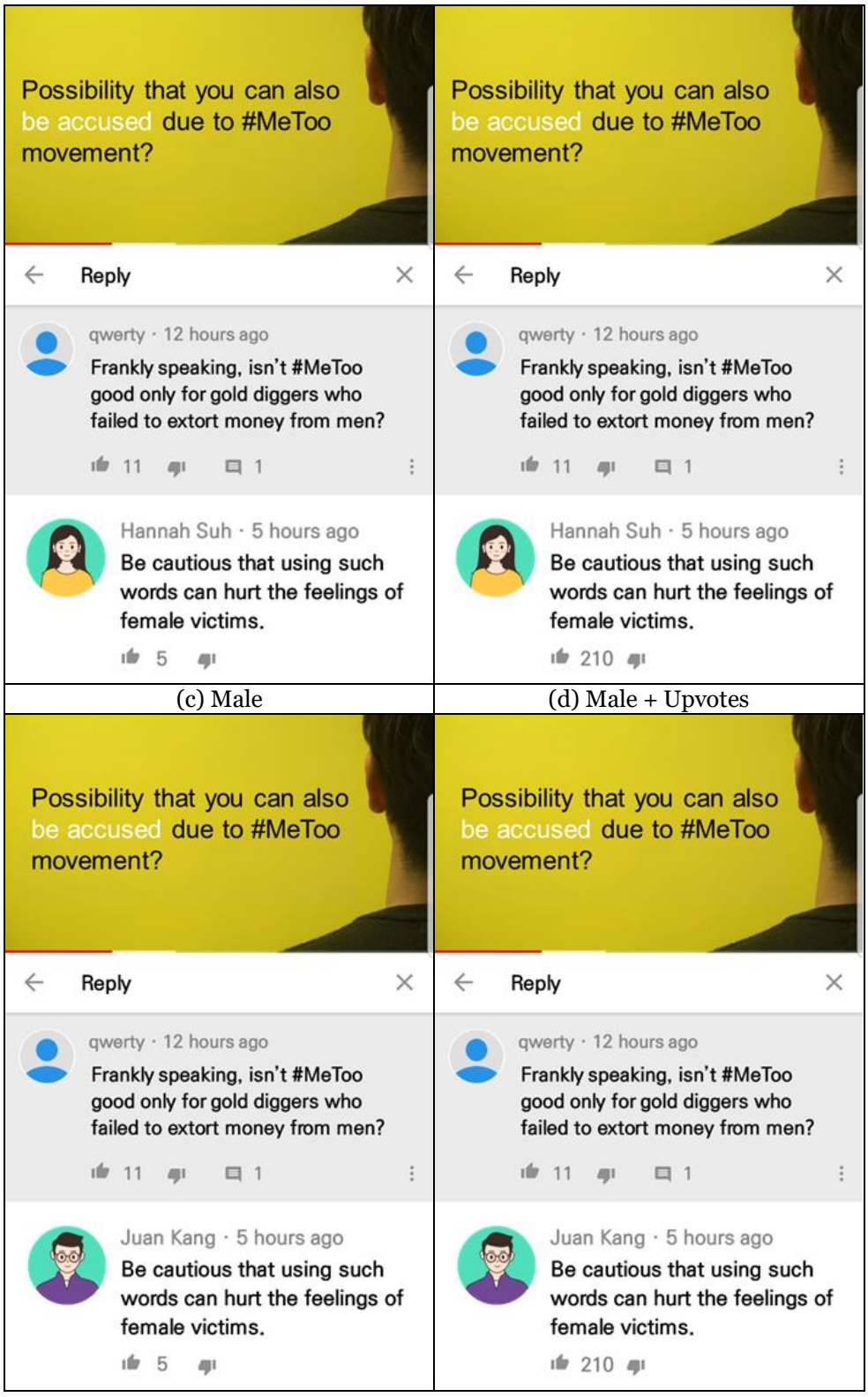

Figure 1. Counter Speech Treatments

To fill this important gap, we examined the research question in the context of sexist hate speech on YouTube via a scenario-based online experiment with a sample of 1,250 adult citizens in South Korea. Focusing on the popularity and the messenger's gender of counter speech, we investigated differential effects of counter speech.

In this experiment, the survey participants were randomly divided into four treatment groups (a 2 by 2 factorial design) and a control group. The survey participants were all exposed to a sexist YouTube comment regarding the Me Too movement that portrays a victim of sexual harassment who came forward during the movement as an opportunist. Participants in the treatment groups were also exposed to another user's counter-speech (reply comment) regarding the sexist YouTube comment. This counter-speech varied in two respects, depending on which treatment condition a respondent was assigned to (a 2 × 2 factorial design): the gender of the replier and the number of upvotes that the reply received. The gender of the replier was manipulated by using gender-specific names and profile images, and the number of upvotes was varied to be either 5 or 210 (see Figure 1). We checked whether the gender and upvote manipulations were properly conducted by querying the participants, during the post-treatment stage of the survey, regarding (a) the gender of the replier (female, male, or "don't know") and (b) the scale of the upvotes the reply received (five-point Likert scale). Statistical analysis showed that participants were able to differentiate the gender manipulation (two-tailed t-test) and upvote manipulation (one-tailed t-test).

Our primary interest is the respondents' intention to report the comment to YouTube, which was measured by a question of "Are you willing to report this comment through the reporting function on YouTube?" with a five-point Likert scale (1 = Very unlikely, 5 = Very likely). In addition, we measured attitude toward the comment by a question of "What do you think about this comment?" (1 = Very negative, 5 = Very positive). Likewise, we measured respondents' attitudes such as attitude toward the commenter, attitude toward the reply, and attitude toward the replier. We estimated the causal effects of these manipulations by calculating the differences between the treatment and control groups (average treatment effect).

The main result showed that both gender and popularity matter when fighting hate speech. We performed t-tests between each treatment group and the control group. In Figure 2, the dots represent group means and the error bars indicate 95% confidence intervals. The respondents were most motivated to report the sexist YouTube comment when the counter speech was unpopular and made by a female user. Of the four treatment conditions, only this condition was statistically significant. When the counter-speech received many upvotes, the female counter-speech effect disappeared. In addition, the treatment and control group respondents did not show statistically significant differences on the other outcome measures—that is, the gender and popularity interventions influenced the respondents' willingness to report the sexist comment but did not affect their views on the commenter, comment, replier, or reply.

We further examined how counter speech affected the respondents' support for regulations on online hate speech. Specifically, we considered two types of regulations: an internal regulation, and an external regulation. In this research, we operationalized the former as YouTube's direct regulation of hate speech and the latter as the French law forcing online platforms to remove within 24 hours hate-speech content (TechCrunch 2020b). For each regulation, we asked respondents the extent to which they support for the regulation using a 5-point Likert scale.

Table 1 shows the results. In columns (1) and (4), the magnitudes were positive and greater for the few-upvotes groups than the many-upvotes groups, these coefficients were statistically insignificant. To understand such differences more clearly, we divided our sample based on the positivity of attitude toward the Me Too movement. We found that the positive effects were positive and significant only for the audience with less positive attitude toward #MeToo. Among

such individuals, the female-authored reply with few upvotes increased the support for the internal (external) regulation by 0.281 (0.250) standard deviation. Although the magnitude is smaller than the female-authored reply, the male-authored reply with few upvotes also significantly increased the support for the regulations for these individuals.

Our research provides important managerial implications. We found that the more upvotes of counter speech discouraged the audience's intention to report hate speech, indicating that the audience may merely support the counter speech and not take direct actions to remove the hate-speech comment from the platform. Online platforms might design messages and interfaces that can emphasize each individual's influence on and responsibility for the online communities to avoid this unexpected consequence. Importantly, the changes in reporting behaviors did not accompany significant changes in the audience's assessment of the hate-speech comment, suggesting that the current data generation process to train machine learning models to predict hate speech might be problematic. To obtain valid labels to detect malicious content, platforms need to quantify and correct the behavioral bias throughout their data collection process.

This study is not without limitations, which could pave ways for future research. First, we tested only two levels of popularity of counter speech. Since the effects of popularity might not be linear, future research may discover when the number of upvotes begins to reduce the audience's proactive actions to counter hate speech. Second, we used a hate-speech comment using an apparent sexist slur. However, one might wonder how counter speech alters an attitude toward controversial speech. Future studies may compare apparent hate speech with controversial one based on a new theoretical ground. Third, our experiment was conducted in a hypothetical setting, which could limit the external validity of our findings. Novel settings that can overcome several practical obstacles, such as observing reporting actions, can significantly contribute to the literature.

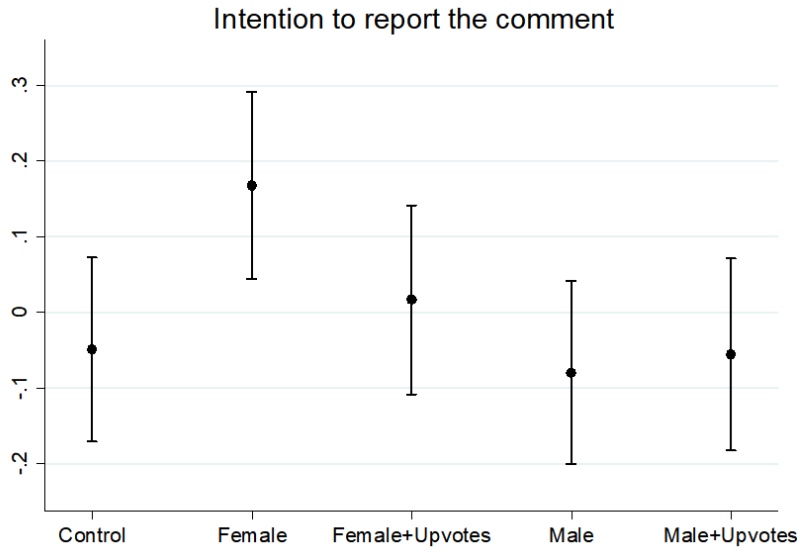

Figure 2. Intention to Report the Comment by Group

Notes. The graph indicates the mean and 95% confidence interval for each group. We use the standardized variable for ease of interpretation.

Dependent Variable:	Attitude toward internal regulation			Attitude toward external regulation		
Respondent Group:	(1) All	Attitude toward #MeToo		(4) All	Attitude toward #MeToo	
		(2) < Median	(3) > Median		(5) < Median	(6) > Median
Treatment (Base: Control)						
Female	0.109	0.281**	-0.0582	0.0773	0.250*	-0.109
	(0.0844)	(0.119)	(0.119)	(0.0831)	(0.129)	(0.102)
Female + Upvotes	0.00607	0.0712	-0.0462	-0.0356	0.191	-0.245**
	(0.0865)	(0.131)	(0.115)	(0.0851)	(0.135)	(0.105)
Male	0.0392	0.204*	-0.0936	0.0304	0.235*	-0.157
	(0.0827)	(0.117)	(0.115)	(0.0833)	(0.128)	(0.108)
Male + Upvotes	0.00199	0.0667	-0.0320	0.0527	0.143	-0.0434
	(0.0857)	(0.133)	(0.110)	(0.0838)	(0.134)	(0.100)
Respondent Controls	Included	Included	Included	Included	Included	Included
Observations	1,250	620	630	1,250	620	630
R-squared	0.121	0.136	0.087	0.149	0.102	0.145

Table 1. Effects of Counter Speech on Support for Hate-Speech Regulations

Notes. Robust standard errors are in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01.

References

Cho, D., and Kwon, K. H. 2015. "The impacts of identity verification and disclosure of social cues on flaming in online user comments," Computers in Human Behavior (51), pp. 363-372.

Döring, N., and Mohseni, M. R. 2020. "Gendered hate speech in YouTube and YouNow comments: Results of two content analyses," Studies in Communication and Media (9:1), pp. 62-88.

Enikolopov, R., Petrova, M., and Sonin, K. 2018. "Social Media and Corruption," American Economic Journal: Applied Economics (10:1), pp. 150-174.

Facebook. 2020. "Community Standards," retrieved from https://www.facebook.com/communitystandards/ (accessed on October 10, 2020).

Fox, J., Cruz, C., Lee, J. Y. (2015) "Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media," Computers in Human Behavior (52), pp. 436-442.

Freelon, D., McIlwain, C., and Clark, M. 2018. "Quantifying the power and consequences of social media protest," New Media & Society (20:3), pp. 990-1011.

Johnson, B. G. 2018. "Tolerating and managing extreme speech on social media," Internet Research (28:5), pp. 1275-1291.

Kwon, K. H., and Cho, D. 2017. "Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites," Social Science Computer Review (35:1), pp. 84-102.

Nakandala, S. C., Ciampaglia, G. L., Su, N. M., and Ahn, Y.-Y. 2017. "Gendered Conversation in a Social Game-Streaming Platform," In Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM), Montreal, Quebec, Canada.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., and Mukherjee, A. 2019. "Thou Shalt Not Hate: Countering Online Hate Speech," In Proceedings

of the International AAAI Conference on Web and Social Media (ICWSM), Munich, Germany, pp. 369-380.
Munger, K. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment," Political Behavior (39:3), pp. 629-649.
Oh, O., Eom, C., and Rao, H. R. 2015. "Role of social Media in Social Change: An analysis of collective sense making during the 2011 Egypt revolution," Information Systems Research (26:1), pp. 210-223.
Shin, J., Jian, L., Driscoll, K., and Bar, F. 2017. "Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction," New Media & Society (19:8), pp. 1214-1235.
Siegel, A. A., and Badaan, V. 2020. "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online," American Political Science Review (114:3), pp. 837-855.
Susarla, A., Oh, J.-H., and Tan, Y. 2012. "Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube," Information Systems Research (23:1), pp. 23-41.
TechCrunch. 2020b. "France passes law forcing online platforms to delete hate-speech content within 24 hours," (May 15) retrieved from https://techcrunch.com/2020/05/14/france-passes-law-forcing-online-platforms-to-delete-hate-speech-content-within-24-hours/
Warren, A. M., Sulaiman, A., and Jaafar, N. I. 2014. "Social media effects on fostering online civic engagement and building citizen trust and trust in institutions," Government Information Quarterly (31), pp. 291-301.
Wu, A. H. 2019. "Gender Bias in Rumors among Professionals: An Identity-based Interpretation," Review of Economics and Statistics, forthcoming.
YouTube. 2020. "Community Guidelines," retrieved from https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ (accessed on October 10, 2020).