Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

PLATFORM GOVERNANCE ON VIOLENCE AGAINST WOMEN: ANALYSIS OF INSTAGRAM, YOUTUBE, TIKTOK AND TWITCH COMMUNITY GUIDELINES

Luiza Carolina dos Santos
Universidade Tecnológica Federal do Paraná (UTFPR); Universidade Federal do Paraná (UFPR)

Raquel Pereira Rodrigues Leite
Universidade Federal do Paraná (UFPR); Pontifícia Universidade Católica do Paraná (PUCPR)

Introduction

In 2024, Safernet Brasil recorded 4,289 complaints of online violence or discrimination against women, making it the third most common complaint that year. When focusing on content aimed at minority groups, violence against women (VAW) accounts for the highest number of reports, a trend since 2018. Women's social vulnerability, online and offline, positions them as primary targets of online hate speech (Lucas, Gomes, Salvador, 2020; Parekh, 2012). They face various violent practices, including doxxing, revenge porn, stalking, and harassment, that can be understood as gender-based violence. The high tolerance for such violence often leads to gendered hate speech being overlooked (Richardson-Self, 2018) and many young women view it as a normal aspect of their online experience (Marwick, Caplan, 2018).

Literature describes violent practices targeting women using terms such as "gendered cyberhate," "gendered e-bile" (Jane, 2017), "cybersexism" (Poland, 2016), and "gendertrolling" (Mantilla, 2013). Additionally, the literature highlights growing phenomena on digital platforms, including gender-based political violence and gendered disinformation, which involve practices like coordinated abuse and gender-based defamation (Judson, 2021). Women in public roles or those expressing feminist views are often targeted for online violence. Yet, this issue continues to be framed as a women’s problem rather than a democratic one (Di Meco, 2019; Soto & Sanchez, 2019). Consequently, many women adopt male personas online, alter their online

Santos, L. C., & Leite, R. P. R. (2025, October). *Platform governance on violence against women: Analysis of Instagram, YouTube, TikTok and Twitch community guidelines*. Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org

consumption and engagement habits, withdraw from the internet altogether (Jane, 2017), and are discouraged from participating in public debate (Sessa, 2020).

Various aspects of the internet contribute to online VAW, including invisibility, instant interactions (Brown, 2018) and digital platforms business models, visibility regimes, and content moderation (Gillespie, 2018). Digital Platforms advocate for minimal regulation and claim neutrality, protecting their market positions while obscuring the lack of autonomy in decision-making through implicit interventions in their user policies (Gillespie, 2010; Mielli, Romanini, 2021). Platform governance relies on terms of service and content guidelines, aiming to encourage positive online behavior while minimizing aggression and antisocial conduct (Flew, Martin, Suzor, 2019).

This paper examines how Twitch, YouTube, Instagram, and TikTok govern gender-based violence in their user policies. We conducted documental research of their Terms of Use, Community Guidelines, and related documents, focusing on the following VAW practices: stalking, doxxing, leaking intimate photos, hate speech, harassment, sexual harassment, gender-based disinformation and political violence.

Method

This work aims to highlight aspects of platform governance regarding VAW, specifically how these platforms propose to self-regulate content within their spaces. We focus on their understanding—or lack thereof—of online VAW, drawing on previous studies that explore platform governance around hate speech (Santos et al., 2023). Through documental research (Sá-Silva, Almeida, Guindani, 2009), we consider that Terms of Use and Community Guidelines shape perceptions about the platform and are aimed at users, stakeholders and governments alike (Gillespie, 2018).

We analyze four platforms with different video formats: Instagram, YouTube, TikTok, and Twitch[1]. Documents selection for analysis began with the Terms of Use, which reference Community Guidelines detailing permissible and impermissible actions on each platform. We included documents linked on their respective pages that were directly relevant to our investigation. Since this paper is part of a broader study on online VAW in Brazil, we incorporated documents in Portuguese, using the English versions only when they were the only ones available. The documents were collected in February 2025 and saved in PDF format due to their frequently updated nature. Our analysis focuses on four aspects: types of restrictions on VAW-related content; how gender appears in the documentation; proposals for combating online VAW; and identification of other vulnerabilities faced by women in these spaces.

Preliminary Results

[1] Instagram, TikTok, and YouTube were among the ten platforms that received the most VAW complaints from Safernet Brasil in 2024.

The platforms studied use a combination of automated and human content moderation, triggered by user complaints and proactive measures[2]. Instagram, YouTube, TikTok, and Twitch prohibit the circulation of hate speech based on protected characteristics (including gender and sex), as well as sexual harassment, non-consensual sharing of intimate images, harassment, bullying, and doxxing. Notably, since January 2025, Instagram has classified certain offenses based on protected categories as 'Level 2,' which are permitted (although not recommended), such as 'claims of mental illness or abnormality when based on gender or sexual orientation, considering political and religious discourse on transgenderism and homosexuality, as well as the common and non-literal use of terms such as "queer"'.

In their disinformation policies, TikTok and Twitch prohibit content that spreads disinformation targeting protected groups (including gender). YouTube and Instagram, however, do not reference protected characteristics in their disinformation guidelines. While TikTok and YouTube restrict political disinformation, neither specifies policies addressing gender-based political violence. TikTok and Instagram adopt a more lenient approach in their harassment and bullying policies, allowing "some negative or critical comments or images" about public figures without acknowledging the heightened vulnerability of women in these roles. Twitch uniquely prohibits stalking, a practice not addressed by the other platforms, and has specific policies regarding harassment or hate speech on "usernames and account display names" and video game content that has been modified by the user.

However, none of the platforms have content moderation policies explicitly designed for online VAW, nor do they propose targeted strategies for tackling it. Policies may cover some types of VAW, but mentions of gender are rare and often presented only as examples. Also, the guidelines don't consider the effects of combined violence practices. Gender appears explicitly as a protected category primarily in hate speech sections. Also, on Twitch and TikTok, participants are encouraged to take responsibility for keeping the platform a safe environment for themselves (in cases of harassment, for example), using tools that make it possible to restrict the visibility of content or forms of interaction.

TikTok maintains a Transparency Center, publishing four annual reports on community guideline enforcement. These reports detail removals by platform policy categories but lack a gender breakdown – the same goes for Instagram and YouTube. TikTok also offers a Safety Center with resources to help users foster a safe environment. Although gender and sexuality are addressed in guides for the LGBTQ+ community, there is no content on vulnerabilities faced by women. Instagram has initiatives aimed at proactively detecting and mitigating the spread of non-consensual sharing of intimate

[2] In January 2025, the founder and CEO of Meta, Mark Zuckerberg, announced changes to content moderation, which are not yet planned to be applied in Brazil but have already appeared in Instagram's guidelines. As a result, moderation will only be applied to legal violations and those considered of high severity, and Instagram will rely on reports. In cases considered of lesser severity, users themselves will be able to add notes (Meta Platforms, 2025).

images and provides guidelines for addressing and reporting it. Twitch, however, lacks information on content removal reports or resources for preventing online violence.

Regarding policies addressing other vulnerabilities faced by women, TikTok, Instagram and YouTube have rules on content about eating disorders and body image. TikTok, restrict this type of content to people over 18 and make it not eligible for recommendation in the ‘For you’ feed. The platform's documents indicate that its system does not make subsequent recommendations for videos about diets ‘to ensure that it is not viewed too often’.

References

Di Meco, L. (2019). Gendered desinformation, fake news, and Women in politics. Council on Foreign Relations. https://www.cfr.org/blog/gendered-disinformation-fake-news-and-women-politics.

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. Journal of Digital Media & Policy, 10(1), 33-50.

Gillespie, T. (2010). The politics of ‘platforms’. New media & society, 12(3), 347-364.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation,and the Hidden Decisions that Shape Social Media. New Haven: YaleUniversity Press. 296 p. ISBN 978-0-30017-313-0.

Instagram. (2025). Transparency Center. https://transparency.meta.com/pt-br/

Instagram. (2017). Not without my consent. https://about.fb.com/wp-content/uploads/2017/03/not-without-my-consent.pdf

Instagram. (2019). Detecting non consensual intimate images. https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/

Jane, E. (2017). Feminist flight and fight responses to gendered cyberhate. In Segrave, M.; Vitis, L. (Eds), Gender, Technology and Violence (pp. 45-61). Routledge: New York.

Judson, E. (2021). Gendered disinformation: 6 reasons why liberal democracies need to respond to this threat. European Union. https://eu.boell.org/en/2021/07/09/gendered-disinformation-6-reasons-why-liberal-democracies-need-respond-threat.

Lemos, A. L. M. (2023). O Futuro da Sociedade de Plataformas no Brasil. Intercom: Revista Brasileira De Ciências Da Comunicação, 46, e2023115.

Luccas, V. N., Gomes, F. V., & Salvador, J. P. F. (2020). Guia de análise de discurso de ódio. Rio de Janeiro: Fundação Getulio Vargas. https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/28626/Guia%20de%20Análise%20de%20Discurso%20de%20Ódio.pdf?sequence=1&isAllowed=y

Mantilla, K. (2013). Genderitrolling: misogyny adapts to new media. Feminist Studies, 39(2), 563-570.

Marwick, A. E., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist media studies*, *18*(4), 543-559.

Meta Platforms. (2025). More Speech and Fewer Mistakes. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes.

Mielli, R., & Vinícius Romanini, A. (2021). A comunicação dominada pelas “big techs” digitais: Superabundância informativa, espetáculo, alienação e fabricação sentido no mundo algorítmico. Revista Eletrônica Internacional De Economia Política Da Informação Da Comunicação E Da Cultura, 23(1), 142–161. https://periodicos.ufs.br/eptic/article/view/14658

Parekh, B. (2012). Is there a case for banning hate speech? In Herz, M., & Molnar, P. (Eds),The Content and Context of Hate Speech: Rethinking Regulation and Responses (pp 37-56). Cambridge: Cambridge University Press.

Poland, B. (2016). Haters: Harassment, Abuse and Violence Online. Lincoln, NB: Potomac Books.

Richardson‐Self, L. (2018). Woman‐Hating: On misogyny, sexism, and hate speech. Hypatia, 33(2), 256-272.

Santos, L. C., Tomaz, R., Dienstbach, D., Matos, E., & Sanches, D. (2023). Discurso de ódio on-line: uma análise das políticas das plataformas digitais para moderação de conteúdo. E-Compós, 26. https://doi.org/10.30962/ec.2709

Sá-Silva, J. R., Almeida, C. D. D., & Guindani, J. F. (2009). Pesquisa documental: pistas teóricas e metodológicas. *Revista brasileira de história & ciências sociais*, *1*(1), 1-15.

Sessa, M. (2022). What is gendered disinformation? Heinrich Böll Foundation, Israel Public Policy Institute (IPPI). https://il.boell.org/en/2022/01/26/what-gendereddisinformation.

Soto, C. A. A., & Sánchez, K. D. V. (2019). Violencia en Internet contra feministas y otras activistas chilenas. Revista Estudos Feministas, 27(3), e58797.

TikTok. (2025). Diretrizes da Comunidade. https://www.tiktok.com/community-guidelines/pt/overview.

TikTok. (2021). An update on our work to safeguard and diversity recomendations. https://newsroom.tiktok.com/en-us/an-update-on-our-work-to-safeguard-and-diversify-recommendations

TikTok. (2025). Centro de Segurança. https://www.tiktok.com/safety/pt-br

TikTok. (2025). Centro de Transparência: https://www.tiktok.com/transparency/pt-br

Twitch. (2025). Diretrizes da comunidade.
https://safety.twitch.tv/s/article/Community-Guidelines?language=pt_BR

Twitch. (2025). Segurança na Twitch.
https://safety.twitch.tv/s/article/Safety-at-Twitch?language=pt_BR

Twitch. (2025). Gerenciando Assédio na Twitch.
https://safety.twitch.tv/s/article/Managing-Harassment?language=pt_BR

YouTube. (2025). Diretrizes da comunidade.
https://www.youtube.com/intl/ALL_br/howyoutubeworks/policies/community-guidelines/#community-guidelines.