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

COMMUNITY-LED MODERATION IN 'THE RUINS' OF TWITTER/X: A CASE STUDY OF THE NORTH ATLANTIC FELLA ORGANIZATION (NAFO)

Community-led moderation in 'the ruins' of platforms

Recent developments in platform governance, such as the disbanding of Twitter/X's Trust and Safety team (Hickey et al., 2025) and Meta's abandonment of professional fact-checking (Watt et al., 2025), point to community-led moderation as an increasingly preferred solution by mainstream social media platforms when dealing with problematic content that does not reach a legal harm threshold, such as disinformation. Research has shown the benefits of involving online communities in content moderation efforts, such as achieving higher effectiveness (Xin et al., 2024) and sensitivity to context (Seering, 2020). As Matias (2019, p. 1) puts it: "Volunteer governance remains a common approach to managing social relations, conflict, and civil liberties online". At the same time, community-led moderation approaches have limitations, such as when volunteer moderators are left to rely on their own heuristics to decide upon complex categories of content (Matamoros-Fernández & Jude, 2025).

Twitter/X offers a salient case study to explore volunteer governance when centralised content moderation fails. Since Musk took over the platform, there has been an increase in the circulation of conspiracy theories, disinformation and hate (Graham & FitzGerald, 2023). While some communities abandoned Twitter/X due to this (DiBenedetto, 2023), others could not afford to do so and continued to 'make a life' among 'the ruins' of the system. In her ethnographic work, Tsing investigates 'the possibility of life' (2015, pp. 227–239) in environments rife with ecological destruction and socioeconomic precarity, what she calls capitalist 'ruins' (pp. 205-214). We appropriate Tsing's metaphor of 'the ruins' to describe what becomes of social media spaces like Twitter/X when they abandon their responsibility to protect their users. Like Tsing, we are interested in investigating communities determined to survive on systems that are crumbling.

The North Atlantic Fella Organization (NAFO) is one such community that remained on Twitter/X's ruins because of the continued importance of the platform in covering Russia's full-scale invasion of Ukraine (Boichak & Kasianenko, in press). Its members, also known as 'fellas', have been active on Twitter/X since May 2022 by debunking and ridiculing online falsehoods spread by highly visible Russian government accounts and pro-Russian actors, reporting problematic behaviour, as well as fundraising on behalf of

Suggested Citation (APA): Kasianenko, K., Matamoros-Fernández, A., & Boichak, O. (2025, October). *Community-led moderation in 'the ruins' of Twitter/X: A case study of the North Atlantic Fella Organization (NAFO).* Paper presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

Ukraine (Kasianenko & Boichak, 2024). However, since late 2022, these users have been reporting decreased effectiveness of the platform's response to hateful conduct.

In this paper, we identify community moderation practices NAFO developed in response to this and interrogate their implications for broader research on community-led moderation in 'the ruins' of platforms.

NAFO's community moderation practices

We draw from 24 semi-structured interviews with members of NAFO, over 4 million tweets mentioning the collective between May 2022 and May 2023, and 189 Twitter/X community notes including the keyword 'NAFO' collected in October 2024. We utilised a grounded theory approach (Corbin & Strauss, 2014) for the analysis of interviews, community notes and a purposive sample of tweets (n=64) which contained mentions of 'community notes', identifying three types of moderation practices relevant to the collective – 'soft-', 'hard-', and 'self'-moderation.

Soft-moderation

Soft-moderation is an approach where problematic content and conduct are not removed from the platform, but have their visibility reduced (Gillespie, 2022) or contextualised, for example, by using a warning label (Zannettou, 2021). For NAFO, there exist three distinct but interconnected ways to engage in soft-moderation. First, while the collective is known for its playful and memetic participation on Twitter/X (Kasianenko & Boichak, 2024), some of the members still see value in publicly debunking false claims (Figure 1).

Figure 1

NAFO fellas have also leveraged Twitter/X's Community Notes to tackle disinformation. We traced the tweets in which they encouraged each other to join the program and rate and write community notes related to the events of the war. Echoing the collective's

understanding of Twitter/X as a battleground in the fight against Russia's disinformation, Community Notes also served as a place for NAFO members to define themselves and their efforts. This was at times achieved through using community notes to ridicule pro-Russian actors portraying NAFO negatively (Figure 2).

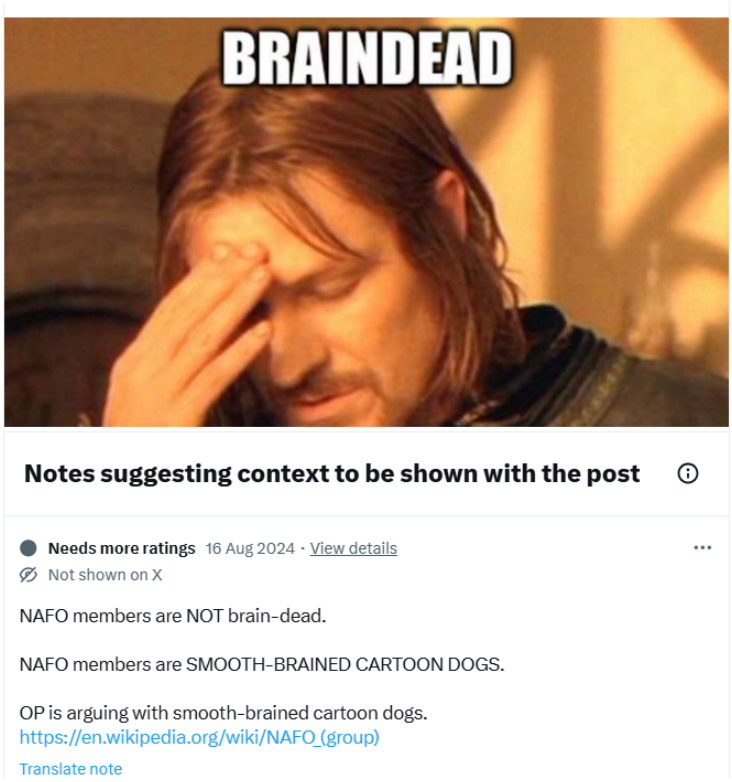

Figure 2

NAFO's imaginative use of community notes is an extension of another common soft-moderation practice led by NAFO fellas on the platform – that of shitposting for the social good. While shitposting refers to the use of humorous or offensive language to derail an online conversation (McEwan, 2017) often for the 'lulz' (Phillips, 2011), NAFO used it to counter Russian propaganda as many members saw engaging with it in good faith ineffective. Instead, they revealed during the interviews that adding memes and jokes under false claims could help highlight their ridiculousness (Figure 3), and at the same provoke the actors spreading such claims and force them to break Twitter/X's policies. This would, in turn, allow 'fellas' to report these actors' content through the platform's official flagging mechanism and potentially have them removed. NAFO members refer to this practice as 'bonking'.

Figure 3

Hard-moderation

NAFO ‘fellas’ who did not want to shitpost engaged in hard-moderation by deliberately and systematically searching for ethnic slurs and common tropes of Russian propaganda on Twitter/X and reporting them. This practice took a considerable emotional toll on the ‘fellas’, especially because of the increasingly perceived ineffectiveness of Twitter/X moderation processes after accounts were reported for violating platforms' policies. As one interviewee put it for us: ‘an account [...] threatening to rape someone or kill someone’ kept coming back ‘in the matter of hours’ of this content being reported and/or removed (Interview 16).

Self-moderation

Faced with the challenge of having to navigate a content moderation system ‘in ruins’, NAFO also increasingly relied on peer support and practices of community work (Kasianenko et al., 2024) to sustain the morale and values of the collective in a platform flooded by Russian propaganda and hate:

> They do it every day, a lot of them[1]. But these are people who in real life are nice people. And they're not used to shit [sic] like this. And so it wears them down. And if I see somebody like that, I'll try to reach out to them. (Interview 23)

Maintaining such a community also meant ensuring the authenticity of new members and fundraising initiatives to guarantee the legitimacy of the collective and its aims, as well as the prevention of in-fighting and helping solve intra-community controversies. The ultimate goal of these self-moderation practices is the reproduction of the collective’s values.

[1] The interviewee here refers to practices of bonking, reporting and driving donation efforts.

Conclusion

Drawing on Tsing's interest in looking at 'possibilities of life' in systems that are 'in ruins', in this paper we have investigated NAFO's perseverance at remaining on Twitter/X despite its moderation system crumbling. By paying specific attention to NAFO's efforts to tackle Russian propaganda and keep their community safe, we identified three important content moderation practices led by members of the organisation: 'soft-', 'hard-', and self-moderation. While, for our participants, Russia's war on Ukraine warranted such efforts, this community-led moderation required a high emotional and time investment on behalf of 'fellas'. Our findings attest to the ability of such practices to fill platform governance gaps (Matias, 2019), while also recognising the need for self-moderation and ongoing care for the community as vital for sustaining such practices.

References

Boichak, O., & Kasianenko, K. (in press). Participatory war and social media: Discursive, material, and ethical dimensions. In A. Bruns, G. Enli, A. O. Larsson, J. Y. Robinson, T. Bosch, & K. Kasianenko (Eds.), *The Routledge Companion to Social Media and Politics* (Second edition). Routledge.

Corbin, J., & Strauss, A. (2014). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.

DiBenedetto, C. (2023, April 21). *LGBTQ centers leave Twitter following removal of hateful conduct protections*. Mashable. https://mashable.com/article/lgbtq-centers-leaving-twitter

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Graham, T., & FitzGerald, K. M. (2023). *Bots, Fake News and Election Conspiracies: Disinformation During the Republican Primary Debate and the Trump Interview*. Digital Media Research Centre, Queensland University of Technology. https://eprints.qut.edu.au/242533/

Hickey, D., Fessler, D. M. T., Lerman, K., & Burghardt, K. (2025). X under Musk's leadership: Substantial hate and no reduction in inauthentic activity. *PLOS ONE*, *20*(2), e0313293. https://doi.org/10.1371/journal.pone.0313293

Kasianenko, K., & Boichak, O. (2024). Canonizing online activism: Memetic iconography in the North Atlantic Fella Organization. *Media, War & Conflict*, 17506352241279957. https://doi.org/10.1177/17506352241279957

Kasianenko, K., Khanehzar, S., Wan, S., Dehghan, E., & Bruns, A. (2024). Detecting Online Community Practices with Large Language Models: A Case Study of

Pro-Ukrainian Publics on Twitter. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 20106–20135). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.1122

Matamoros-Fernández, A., & Jude, N. A. (2025). The importance of centering harm in data infrastructures for 'soft moderation': X's Community Notes as a case study. *New Media and Society*. https://eprints.qut.edu.au/254907/

Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, *5*(2), 2056305119836778. https://doi.org/10.1177/2056305119836778

McEwan, S. (2017). Nation of shitposters: Ironic engagement with the Facebook posts of Shannon Noll as reconfiguration of an Australian national identity. *PLATFORM: Journal of Media & Communication*, *8*(2).

Phillips, W. (2011). LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*. https://doi.org/10.5210/fm.v16i12.3168

Seering, J. (2020). Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW2), 107:1-107:28. https://doi.org/10.1145/3415178

Tsing, A. L. (2015). *Mushroom at the End of the World: On the Possibility of Life in Capitalist Ruins*. Princeton University Press.

Watt, N., Riedlinger, M., & Montaña-Niño, S. (2025, January 8). *Meta is abandoning fact checking – this doesn't bode well for the fight against misinformation*. The Conversation. http://theconversation.com/meta-is-abandoning-fact-checking-this-doesnt-bode-well-for-the-fight-against-misinformation-246878

Xin, W., Wang, K., Fu, Z., & Zhou, L. (2024). *Let Community Rules Be Reflected in Online Content Moderation* (arXiv:2408.12035). arXiv. https://doi.org/10.48550/arXiv.2408.12035

Zannettou, S. (2021). "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 865–876. https://doi.org/10.1609/icwsm.v15i1.18110