Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

DOES ALGORITHMIC CONTENT MODERATION RPOMOTE DEMOCRATIC DISCOURSE? RADICAL DEMOCRATIC CRITIQUE OF TOXIC LANGUAGE AI

Dayei Oh
University of Helsinki

John Downey
Loughborough University

Introduction:
Platforms play a significant role in shaping political discourse, yet concerns persist regarding the proliferation of harmful content jeopardising the inclusivity and safety of online spaces. In response, platforms have increasingly turned to algorithmic content moderation, employing machine learning AI to detect and manage potentially toxic material (Dixon et al.,2018; Gorwa et al., 2020). Google's Perspective API is an example of toxic language moderation algorithms to remove content that are 'rude, disrespectful, or unreasonable […] that is likely to make someone leave a discussion' (Dixon et al., 2018: 68). However, critiques abound regarding the efficacy and biases inherent in such systems, prompting a re-evaluation of their role in fostering democratic discourse.

There have been numerous critiques of the technical limitations of algorithmic moderation including its susceptibility to biases and misclassification (Gorwa et al., 2020; Thiago et al., 2020; Zhou, 2021). AI communities are researching ways to mitigate such biases by experimenting with new ways to deal with these problems framed as 'technical glitches' – incidental errors that can be patched up through machinery itself: e.g., using alternative methods for dividing training and testing datasets, mitigating spurious correlations with additional mathematical methods (Zhou, 2021).

However, the fundamental problems lie much deeper than the incidental 'glitches.' Rather, the problems are embedded in the very logic of content moderation that certain forms of language are 'toxic' and must be censored to promote democratic discourse. The problem grows as soon as the tech communities' definition of 'toxic language' is

Suggested Citation (APA): Oh, D., Downey, J. (2024, October). *Does Algorithmic Content Moderation Rpomote Democratic Discourse? Radical Democratic Critique of Toxic Language Ai*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

loosely and inconsistently defined without critical engagements with theories of democracy, democratic discourse, civility, and tolerance. This paper provides a radical democratic critique of algorithmic moderation both from normative theories of inclusive public spheres and empirical evidence.

Against regulating incivility: Incivility is not tantamount toxic language

The first two key components of 'toxic language' in Perspective API is 'rude' and 'disrespectful' language. At first, equating incivility to anti-democratic toxic language sounds unproblematic, given that civility is seen as an essential component of liberal condition. In pluralist society in which it is difficult for its people to agree on every aspect of values and aspects, civility as a form of politeness and self-control acts as a sort of social glue, encouraging us to believe despite disagreements that we belong to the same political community. Rawls (1996) argues for a duty of civility based on the idea of reciprocity and the practice of public reason, listening to others, fairmindedness, and making reasonable accommodations to the views of others (p.217).

However, there is another diametrically opposed tradition which sees civility as deeply exclusionary. Elias (2000) argues that civility has become a way of excluding people. The conventions of civility are decided by the elites and used to legitimise the exclusion of the marginalised who are seen as morally inferior or non-conforming to the elite's rule of civility. Feminist scholars make similar connection regarding the norm of civility as an exclusionary criterion largely applied to the voices of the marginalised as well as sabotaging the radical social changes that can successfully fight against existing inequalities and injustice (Bickford, 2011; Meyers, 2009; Zerilli, 2014). From the radical democratic perspectives, uncivil political actions extend democracy with expressive and instrumental values under the conditions of existing participatory inequalities, not hampers it (Edyvane, 2020; Young, 2000). For toxic language AI to moderate rude, disrespectful, and uncivil language as 'toxic' is to miss these important roles of incivility.

For regulating (intolerance) and hate speech

But then what kind of content should be moderated? Arguing on the same radical grounds to see the virtue of democracy and democratic discourse not as a liberal notion of civility but as promotion of participatory equality between the majority and minorities, we argue that algorithmic moderation should focus on detecting intolerance and hate speech. This conception of toxicity aligns better with many platforms' promises to tackle hate content as well as 'borderline content' (YouTube, 2019).

If (in)civility is to do with a speech style, (in)tolerance is to do with moral-political attitudes to others – seeing the other side and the minorities as citizens with equal political status which must be included in the public discourse (Forst, 2003; Marcuse, 1969). When intolerant content is directly targeted at groups and members who are subjugated to systemic discrimination, the harm of such intolerant speech amounts to hate speech in terms of its constitutive and consequential harm, being sufficient to warrant moderation (Gelber, 2021).

Separating incivility and intolerance is particularly important since not every hate speech relies on explicit hateful expressions, slurs, and extreme emotions (Gelber, 2021). Previous research discusses borderline discourse of far-right actors online and how

they communicate intolerant messages with quasi-academic, pseudo-rational civil language, and humour (Krzyżanowski & Ledin, 2017; Thiago et al., 2020)

Empirical evaluation of toxic language AI

We then empirically assess the efficiency of Perspective API in detecting incivility and intolerance on large Twitter datasets collected during abortion constitution discussions in Ireland (2018, 1.8+ million tweets) and the US (2020, 6+ million tweets). We use lexicon-based classification of incivility and intolerance, automatic gender recognition, and abortion issue stance mining to assess whether Perspective's understanding of toxicity meets our critical theory-driven understanding of democratic discourse.

First, we find that Perspective's understanding of toxicity is biased towards detecting incivility and not intolerance, which is counterproductive to deepen democratic discourse as we theorised earlier that incivility carries significant expressive and instrumental values in democracy to expand the participatory equalities for the marginalised (Edyvane, 2020; Young, 2000; Zerilli, 2014). Second, by equating toxicity to incivility, Perspective gives higher toxicity scores to tweets written by women and pro-abortion rights users in both countries. Like many feminist and critical theorists of democracy have argued, anger and incivility are tools for those who fight for equality under the conditions of injustice and moderating incivility largely leads to disproportionate silencing of the marginalised and minorities who do not wish to and who are not able to behave civilly to their oppression and discrimination (Bickford, 2011; Meyers, 2009; Zerilli, 2014). Third, we find that Perspective's detection of intolerance and hate speech depends largely on the presence of explicit slurs and hateful terms, while missing out many intolerant tweets (e.g., misogynistic, transphobic, and homophobic) whose hateful ideas are embedded in nuances and rhetoric of seemingly civil, quasi-intellectual language, humour, and satire (Krzyżanowski & Ledin, 2017; Thiago et al., 2020).

Concluding discussion: Future of algorithmic moderation

Based on our normative and empirical critiques, we argue that the current toxic language moderation algorithms do not promote democratic discourse but hinder it. Future algorithmc moderation should focus on detecting intolerance and hate speech, and not incivility. Algorithmic moderation with context awareness is required to take into account the identity of the speakers when assessing toxicity of political opinion expressions (e.g., slurs used by LGBTQ+ to reclaim the slur vs. used to spread hate; Thiago et al., 2020).

Consequently, we recommend that the platforms consider new algorithmic moderation developed in close collaboration with the theories of democracy and democratic public spheres informed by anti-racist, feminist, and other critical theorists. We also recommend that the development of algorithmic moderation should focus on the reliable and transparent identification of intolerance and hate speech for its mission to tackle moderating borderline content.

Note:
The full manuscript of this extended abstract is published open access. If you wish to cite this work, please use the following citation:

Oh, D., & Downey, J. (2024). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. *Information, Communication & Society*, 1–20. DOI: https://doi.org/10.1080/1369118X.2024.2346531

References

Bickford, S. (2011). Emotion talk and political judgment. *The Journal of Politics*, 73(4), 1025-1037.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Edyvane, D. (2020). Incivility as Dissent. *Political Studies*, 68(1), 93–109. https://doi.org/10.1177/0032321719831983

Elias, N. (2000). *The Civilizing Process: Sociogenetic and Psychogenetic Investigations,* Revised edn. Oxford: Blackwell.

Forst, R. (2003). Toleration, justice and reason. In C. McKinnon & D. Castiglione (Eds.), *The culture of toleration in diverse societies* (71–85). Manchester University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society,* 7(1), 2053951719897945.

Gelber, K. (2021) Differentiating hate speech: a systemic discrimination approach, *Critical Review of International Social and Political Philosophy*, 24:4, 393-414, DOI: 10.1080/13698230.2019.1576006

Krzyżanowski, M., & Ledin, P. (2017). Uncivility on the web. *Journal of Language and Politics*, 16(4), 566–581. https://doi.org/10.1075/jlp.17028.krz

Marcuse, H. (1969) Repressive tolerance. In R. P. Wolff, B. Moore Jr, & H. Marcuse (Eds.), *A critique of pure tolerance* (93-138). Jonathan Cape.

Meyers, D. T. (2018). *Feminists rethink the self*. Routledge.

Rawls, J. (1996). *Political Liberalism*. Cambridge, MA: Harvard University Press.

Thiago, D. O., Marcelo, A. D., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & culture*, 25(2), 700-732.

Young, I. M. (2000). *Inclusion and democracy*. Oxford University Press.

Zerilli, L. M. G. (2014). *Against civility: A feminist perspective. Civility, Legality, and Justice in America*, Winter 2010, 107–131. https://doi.org/10.1017/CBO9781107479852.005

Zhou, X. (2021). *Challenges in automated debiasing for toxic language detection*. University of Washington.