Selected Papers of #AoIR2020:
The 21st Annual Conference of the
Association of Internet Researchers
Virtual Event / 27-31 October 2020

TUNING OUT HATE SPEECH ON REDDIT: AUTOMATING MODERATION AND DETECTING TOXICITY IN THE MANOSPHERE

Dr Verity Trott
Monash University

Dr Jennifer Beckett
University of Melbourne

Venessa Paech
University of Sydney

Tuning out hate speech on Reddit: Automating moderation and detecting toxicity in the Manosphere

Over the past few years social media platforms have been struggling to moderate at scale. At the same time, they have come under fire for failing to mitigate the risks of perceived 'toxic' content or behaviour on their platforms. Discussion has turned to the role that automated machine-learning (ML) tools might play in an effort to better cope with content moderation, to combat hate speech, 'dangerous organisations' and other bad actors present on platforms. In 2017, Jigsaw announced a technological answer to the problem of hate speech: a machine-learning technology called *Perspective* that promises to detect toxicity in online comments with the goal of automating moderation. In late 2019, Reddit, marketed as the 'front page of the internet,' announced that they would be using automated technology like *Perspective* to help moderate harassment and bullying on their platform.

This paper contributes to thinking about the role and suitability of ML for content moderation on community platforms such as Reddit and Facebook. In particular, it looks at how ML tools operate (or fail to operate) effectively at the intersection between online sentiment within communities and social and platform expectations of acceptable discourse. Through an examination of the r/MGTOW subreddit we problematise current understandings of the notion of 'toxicity' as applied to cultural or social sub-communities, like MGTOW, online and explain how this interacts with Google's Perspective tool.

Suggested Citation (APA): Trott, V., Beckett, J., Paech, V. (2020, October). *Tuning out hate speech on Reddit: Automating moderation and detecting toxicity in the Manosphere.* Paper presented at AoIR 2020: The 21th Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

Reddit has found itself host to a range of communities that are connected to the Manosphere, a loose confederacy of male-only groups, that have gained public attention for perpetuating violent and misogynistic attitudes (Ging 2017: 638). Some of these groups have been directly tied to incidences of mass violence offline provoking media scrutiny into the role of platforms like Reddit in hosting violent men's groups (Kalish and Kimmel 2010; Nicholas and Agius 2018). Reddit has responded to the mounting public scrutiny by sanctioning subreddit communities that are deemed offensive to the mainstream public. Further, they have updated their harassment and bullying policies and announced the enhancement of their moderation efforts with the employment of machine-learning technology.

The subreddit for MGTOW (also known as Men Going Their Own Way) was chosen as a site for examination because of their status as the largest growing community belonging to the Manosphere (a loose confederacy of men's groups that have garnered attention for their role in several violent mass physical attacks) (Jones, Trott, Wright 2019). The top ten threads of all time on r/MGTOW were scraped in October 2019 using the Reddit API and included every comment within these threads, resulting in a dataset of 2,861 comments. These comments were initially analysed for toxicity using Perspective.

Comments were rated as either not toxic, 'quiet,' 'low,' 'moderate,' 'loud,' and 'blaring'. It is not clear how these categories of toxicity are defined but the developer information for *Perspective* states that comments are scored based on 'the perceived impact a comment might have on a conversation,' (Jigsaw, 2019). In other words, a comment's discursive function, rather than the attitudes and beliefs depicted within the comment. Overall 70% of comments were filtered within r/MGTOW but the majority of comments were rated as 'low' toxicity (30%) and less than 10% were considered 'loud' and only 1% were labelled as 'blaring' in toxicity. While the overall results registered low levels of toxicity within the community it has become clear that Reddit does not view the group in the same light. As of 6th February 2020, r/MGTOW has been quarantined by reddit for breaching the platform's community guidelines.

To interrogate the discrepancy between Perspective's understanding of 'toxicity' and that of Reddit, further analysis was conducted on the content of the subreddit. This involved deductive coding of the content based on Reddit's internal community guidelines and the *Sense of Community Index-2 (SCI-2)*. The SCI-2 measures community solidarity (sense of community) across four broad categories: membership, influence, reciprocity and shared emotional connection and is used by community management professionals and sociologists as a measure of community 'health'.

Our overall analysis points to a tension between current social framings and operationalised notions of 'toxicity' and sociological understandings of community health as framed by the *SCI-2*. Understanding these tensions exposes issues around the use of ML tools for the automated moderation of community spaces within platforms. Of particular interest is the finding that automated ML tools, especially those that rely on sentiment analysis as Perspective does, may be good tools for measuring individual community health, making them potentially useful to community management

practitioners. Community health in this sense, however, is not what we consider when we talk about the effects of groups such as MGTOW within the broader fabric of society.

Previous scholars (Papacharissi 2002) have conceptualized online spaces such as Reddit to be an extension of the public sphere thus it is important to redefine 'toxicity' to take into account more than corporate interests that frame toxicity around platform engagement. The key contribution of this paper is theorizing a framework for understanding and defining 'toxicity' in moderation in terms of different levels: corporate, community-specific, and broader societal interests. It draws attention to the different logics at play that collide and intersect in moderation processes, which have resulted in several challenges and key failings in the development of automated moderation tools and the defining of 'toxicity'.

References

Ging, D. (2017). Alphas, betas, and incels: Theorising the masculinities of the manosphere. *Men and Masculinities*, 22(4): 638-657.

Jigsaw. (2019) *PerspectiveAPI,* website, viewed 22 November 2019. Available at: https://www.perspectiveapi.com/#/home

Jones, C., Trott, V., and Wright, S. (2019) Sluts and Soyboys: MGTOW and the production of misogynistic online harassment. *New Media & Society*. Epub ahead of print 8 November. DOI: 10.1177/1461444819887141.

Kalish, R. and Kimmel, M. (2010). Suicide by mass murder: masculinities, aggrieved entitlement, and rampage school shootings. *Health Sociology Review* 19(4): 451-464.

Nicholas, L. and Agius, C. (2018). #Notallmen, #Menenism, Manospheres and Unsafe Spaces: Overt and Subtle Masculinism in Anti-"PC" Discourse. (pp. 31-59) In: Nicholas L and Agius C (eds) *The Persistence of Global Masculinities* New York: Springer.