Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

MODERATING MENTAL HEALTH: ARE AUTOMATED SYSTEMS TOO RISK AVERSE?

Anthony McCosker
Swinburne University of Technology, Melbourne, Australia

Introduction

Across commercial social media platforms and dedicated support forums alike, mental health content raises important questions about what constitutes risk and harm online and how to moderate it. Mental ill-health is treated by most social media moderation systems as 'borderline' or problematic content, with access and posts often restricted to reduce assumed secondary harms (McCosker & Gerrard, 2021). This paper explores how automated and human moderation practices can be re-configured to accommodate resilient behaviours and social support.

Drawing on work with three Australian mental health organisations that provide successful discussion and support forums, I identify human and machine moderation practices that can help to re-think how mental health content is managed. The work aims to improve safety and resilience in these spaces, drawing insights from successful practices to inform the treatment of mental health content more widely across social media. Through an analysis of interviews and workshops with forum managers and moderators, I argue that platforms must incorporate strengths-based context (resilience indicators) into their moderation systems and practices, challenging simplistic assessments of mental health content as risk and harm.

Background

Gillespie and Aufderheide (2020, p2) define content moderation in relation to *unacceptable* content and behaviour. However, mental health content poses different challenges to misinformation or race hate speech, for example. This content is often deemed unacceptable through its capacity (whether real or imagined) to trigger others and encourage self-harm or suicidality. For instance, in her account of content moderation, Gerrard poses the question of whether Instagram should allow currently

Suggested Citation (APA): McCosker, Anthony. (2022, November). *Moderating Mental Health: Are Automated Systems Too Risk Averse?* Paper (or panel) presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

banned images of healed self-cutting, under the guise of supporting recovery (2022, p. 86-87). Facebook and Instagram have received sustained criticism for their effect particularly young women's mental health, but this is not linked directly to content dealing with mental health, or associated aesthetics and memetics (McCosker & Gerrard, 2021). I argue that insights can be drawn from those who undertake moderation of dedicated mental health support forums to address these issues on a broader scale, while improving the professionalisation of content moderation targeted at expressions of mental (ill)health.

Specifically, I argue for the need to better balance moderation practices (automated and human) within a framework of strengths-based indicators of resilience and a logic of care (Mol, 2008). The concept of resilience used here is drawn from literature that spans both individual and community-level factors. These involve resilience generators: learning, social capital and belonging; and resilience characteristics: adaptive capacity and self-efficacy (Berkes and Ross, 2013). These indicators of resilience are being developed and tested through an associated project and were incorporated into discussions and interviews with forum managers and moderators (see Kang et al., 2022).

Methods

This paper focuses on qualitative research with seven forum moderators and managers from three nonprofit organisations, providing analysis of discussions, in-depth interviews, in the context of moderation policy and community guidelines. Initial analysis of these interviews draws common themes focusing on interactions between automated and human moderation processes, contextual cues regarding high-level goals and guidelines for moderation, and 'adaptive practices' – or the points where moderators have had to realign their moderation practices in support of forum members.

Each work on the Forum platform provided by Khoros, with small teams of 24/7 moderation teams and member moderators (people with lived experience who act as volunteer moderators). These platforms seek to scale and improve their moderation practices and outcomes. Each also use a bespoke automated triage system, and my analysis targets the interaction between these systems and the human moderator teams and their everyday moderation practices.

Adaptive moderation practices and the human-AI alignment problem

Forum managers and moderators refer to their overall goals explicitly in terms of resilience and see their services as offering a unique form of preventative care. Drawing on common internet and social media metaphors, they refer to the forums, sometimes in direct contrast with Facebook and other platforms, as a 'safe space' (Org 1). The goal of 'building self-efficacy' and enabling members to 'learn from others' is seen as a pre-cursor to developing 'adaptability' (Org 2). By increasing their members' 'self-advocacy', 'their knowledge and their readiness to take a more intensive step', moderators see forums as enablers of resilience and recovery (Org 3). Given these high-level goals, it

was notable that moderators and managers focus their daily processes on addressing ‘risky’ content identified by the automated triage systems.

As with commercial platforms, moderation practices at these organisations are layered. An automation layer built on a supervised machine learning classifier is trained to rank messages as green, amber, red or crisis depending on the urgency of attention required. However, there is evidence of moderator teams applying *adaptive* moderation practices, accommodating members in a more *ad hoc* and responsive fashion. One participant talked about the way their team’s responses to flagged content are based on training and intuition, as well as through use of a quick reference guide (Org 3). The team draws on their experience, but ‘we also have a look at the reference guide in terms of *how* to respond, like what are we looking for, or what young people are looking for when we're responding’ (org 3). Another noted that: ‘We have had to adapt with Covid to allow people to use the forums as a bit of a vent space…people are frustrated, and they're allowed to be frustrated.’ (Org 2).

Previous research has incorporated other contextual factors to help interpret and classify posts through an automated moderation system. This includes attributes from: ‘content and structure of the user history, other posts in the conversation and the interaction network’ (Altszyler et al., 2018: 63). There is room to explore this contextual layer further by incorporating, for instance, resilience indicators and attributes in both the automated classifier models, and manual moderation practices. This would constitute an *adaptive moderation* system; one that can incorporate context, events, and other factors such as place.

When ‘integrity’ and content moderation systems are trained primarily to find and restrict content designated as risky and harmful, they provide an incomplete response to mental health interaction and content. Both the automated and human moderators need to weigh these risks and harms against strengths - what I have referred to in this paper as resilience factors.

References

Altszyler, E., Berenstein, A. J., Milne, D. N., Calvo, R. A., & Slezak, D. F. (2018, June). Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 57-68).

Berkes, F., and Ross, H. (2013). "Community resilience: toward an integrated approach." *Society & Natural Resources* 26 (1):5-20.

Gillespie, T., and Aufderheide, P. (2020). Introduction: Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), 1–29.

Girrard, Y. (2022). What is content moderation? In D. Rosen (Ed.). *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media*. Routledge, 77-95.

Kang, Y. B., McCosker, A., Kamstra, P., & Farmer, J. (2022). Resilience in web-based mental health communities: Building a resilience dictionary with semiautomatic text analysis. *JMIR formative research*, *6*(9), e39013.

McCosker, A., & Gerrard, Y. (2021). Hashtagging depression on Instagram: Towards a more inclusive mental health research methodology. *New Media & Society*, *23*(7), 1899-1919.

Mol, A. (2008). *The logic of care: Health and the problem of patient choice*. Routledge.