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

MEASURING MISOGYNY: *DEPP V HEARD* AND THE LIMITS OF ATOMISTIC CONTENT MODERATION

Lucinda Nelson
Queensland University of Technology

Nicolas Suzor
Queensland University of Technology

Introduction and background

There is a fundamental disconnect between existing approaches to content moderation and the dynamics of violence against women. In this paper, through an examination of the online discourse around the *Depp v Heard* trial, we illustrate the limitations of atomistic content moderation. We argue that online misogyny cannot be adequately addressed through the moderation of individual pieces of content, and that we need new methods for measuring misogyny in aggregate. This paper is part of a larger, ongoing project that seeks to understand how to better identify and address the persistent undercurrent of 'everyday' misogyny on social media platforms. We use 'everyday misogyny' to refer to the subtler expressions of online misogyny that do not reach the threshold for prohibition or removal under either the law or platform policies, but still contribute to broader platform cultures that are hostile to women.

Over the past decade, social media companies have come under increasing pressure to make their platforms safer for women. While they have made some changes – banning specific groups and pages; committing to reducing the spread of "borderline content"; introducing features that allow users to add additional context when reporting content – the foundational approach to content moderation has remained largely the same. Platforms continue to focus on identifying, assessing and responding to individual pieces of violating content, like overt hate speech, direct threats, and doxxing. This approach is ill-equipped to deal with structural harms, like misogyny (Suzor 2019). Scholars have emphasised that violence against women cannot be understood in terms of isolated instances, but as part of a continuum (Kelly 1987). 'Everyday' experiences of sexism and misogyny form part of the same dynamic as the more widely recognised,

Suggested Citation (APA): Nelson, L and Suzor, N. (2024, October). *Measuring misogyny: Depp v Heard and the limits of atomistic content moderation*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

extreme forms of violence (Gillett 2018). This makes everyday misogyny an important, albeit controversial, site for platform intervention.

The most well-known existing tools for identifying harmful content tend to focus on variations of 'toxic' content, 'not safe for work' material, or explicit hate speech. For example, Perspective API, the current industry standard for measuring 'toxicity', seeks to identify comments that are "rude, disrespectful or, unreasonable" and are "likely to make someone leave a discussion". Because they do not deal with context, these methods of assessing short texts in isolation often tend to function primarily as a detector of incivility (Trott, Beckett, and Paech 2022). As a result, tools like Perspective have been shown to tone-police the content of marginalised users, while also failing to identify politely worded expressions of harmful ideologies (Dias Oliva, Antonialli, and Gomes 2021).

Methods

This paper explores these current tensions in content moderation using the case study of the *Depp v Heard* trial. The online response to this defamation trial, brought by Johnny Depp against Amber Heard, was emblematic of the anti-feminist backlash against the #MeToo movement. While there was no shortage of overtly hateful content in the online discourse, this case study also presents an interesting opportunity to explore everyday forms of misogyny. The combination of the celebrity element and the public broadcasting of the trial created a spectacle, generating widespread discussion about the trial and allowing online spectators to participate in the interpretation and evaluation of evidence. As a legal trial, where the 'truth' was inherently contested, this case study is particularly useful for investigating perceptions of believability (Banet-Weiser and Higgins 2023) and doubt in the context of violence against women – concepts that are both deeply influenced by harmful beliefs about women, and widely seen as legitimate points for open discussion and debate.

This data for this study is a set of over two million unique tweets containing relevant keywords, including variations and combinations of 'johnny depp', 'amber heard', 'depp v heard', posted between 4 April 2022 (one week before the trial began) and 8 June 2022 (one week after the trial ended). The first stage of our methods was exploratory. We developed a thorough understanding of the timeline of the trial and key issues in the discourse by examining peaks of activity in the data, popular hashtags and links, and influential accounts with tweets that were widely shared or replied to. We then used topic modelling to organise the data into identifiable topics of interest, focusing on topics related to evidence, believability and doubt. With the data organised by topic, we undertook a closer qualitative analysis of a random sample of each of our topics of interest. This analysis was informed by the literature on believability and doubt in domestic and sexual violence cases, including testimonial injustice (Fricker 2007; Harradine 2022; Hänel 2022), rape and domestic violence myths (Burt 1980; Stabile et al. 2019; Peters 2008; Policastro and Payne 2013), and 'ideal victim' narratives (Christie 1986; Randall 2004).

Preliminary results

This paper is a work in progress. Our preliminary findings indicate that everyday misogyny was widespread in online discussions about the trial. Posts about evidence were often grounded in harmful beliefs about women and about the nature of domestic and sexual violence, including that women frequently lie about experiencing abuse and that victim-survivors will always act in a particular way. We also found significant double-standards in what was expected of each party, both in terms of their behaviour in court and the type and amount of evidence needed to prove their claims. Users also expressed contradictory expectations of Amber Heard, which culminated in a series of double-binds: she was criticised for crying too much, but also for smiling or laughing; she was criticised for not having enough evidence of her physical injuries, but the fact that she recorded instances of Depp's abusive behaviour was also perceived as suspicious. Importantly, these double-standards and double-binds were rarely found within the same post, or even the same comment thread. The full picture of the misogyny in the online discourse only becomes apparent through consideration of the culmination of tweets. Our analysis suggests that it is not enough to perform a binary misogynistic/not misogynistic classification at scale; rather, we need to consider the broader patterns, recognising that entirely separate pieces of content still contribute to a common discourse and have cumulative weight.

Contribution

The central contribution of this paper is to provide deeper insights into the way everyday misogyny manifests online, and the limitations of current approaches to content moderation. We intend to build on the findings of this study to develop new methods and frameworks for understanding misogyny in aggregate. Through this work, we aim to inform ongoing debates in the content moderation literature about what platforms should do to address structural harms.

References

Banet-Weiser, Sarah, and Kathryn Claire Higgins. 2023. *Believability: Sexual Violence, Media, and the Politics of Doubt*. 1st ed. Polity.

Burt, M. R. 1980. 'Cultural Myths and Supports for Rape'. *Journal of Personality and Social Psychology* 38 (2): 217–30. https://doi.org/10.1037//0022-3514.38.2.217.

Christie, Nils. 1986. 'The Ideal Victim'. In *From Crime Policy to Victim Policy: Reorienting the Justice System*, edited by Ezzat A. Fattah, 17–30. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-08305-3_2.

Dias Oliva, Thiago, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online'. *Sexuality & Culture* 25 (2): 700–732. https://doi.org/10.1007/s12119-020-09790-w.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Clarendon Press.

Gillett, Rosalie. 2018. 'Intimate Intrusions Online: Studying the Normalisation of Abuse in Dating Apps'. *Women's Studies International Forum* 69 (July): 212–19. https://doi.org/10.1016/j.wsif.2018.04.005.

Hänel, Hilkje C. 2022. '#MeToo and Testimonial Injustice: An Investigation of Moral and Conceptual Knowledge'. *Philosophy & Social Criticism* 48 (6): 833–59. https://doi.org/10.1177/01914537211017578.

Harradine, Michelle. 2022. 'Defamation Law and Epistemic Harm in the #MeToo Era'. *Australian Feminist Law Journal* 48 (1): 31–55. https://doi.org/10.1080/13200968.2022.2146303.

Kelly, Liz. 1987. 'The Continuum of Sexual Violence'. In *Women, Violence and Social Control*, edited by Jalma Hanmer and Mary Maynard, 46–60. Palgrave Macmillan.

Peters, Jay. 2008. 'Measuring Myths about Domestic Violence: Development and Initial Validation of the Domestic Violence Myth Acceptance Scale'. *Journal of Aggression, Maltreatment & Trauma* 16 (1): 1–21. https://doi.org/10.1080/10926770801917780.

Policastro, Christina, and Brian K. Payne. 2013. 'The Blameworthy Victim: Domestic Violence Myths and the Criminalization of Victimhood'. *Journal of Aggression, Maltreatment & Trauma* 22 (4): 329–47. https://doi.org/10.1080/10926771.2013.775985.

Randall, Melanie. 2004. 'Domestic Violence and the Construction of Ideal Victims: Assaulted Women's Image Problems in Law Deconstructing the Image of Battered Woman'. *Saint Louis University Public Law Review* 23 (1): 107–54.

Stabile, Bonnie, Aubrey Grant, Hemant Purohit, and Mohammad Rama. 2019. '"She Lied": Social Construction, Rape Myth Prevalence in Social Media, and Sexual Assault Policy'. *Sexuality, Gender & Policy* 2 (2): 80–96. https://doi.org/10.1002/sgp2.12011.

Suzor, Nicolas. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

Trott, Verity, Jennifer Beckett, and Venessa Paech. 2022. 'Operationalising "Toxicity" in the Manosphere: Automation, Platform Governance and Community Health'. *Convergence*, June, 1–16. https://doi.org/10.1177/13548565221111075.