RESEARCHBRIEF
No.2 JULY, 2026

Defining and Detecting Online Misogyny: Aligning Social and Computer Sciences

Isha Mangurkar, Visiting Research Fellow, United Nations University Institute in Macau, Macau SAR, China
https://orcid.org/0000-0001-7052-5644

Jaimee Stuart, Head of Research (Acting), United Nations University Institute in Macau, Macau SAR, China
https://orcid.org/0000-0002-4376-1913

The rapid adoption of Information and Communication Technologies (ICTs) has made the world more interconnected than ever. This shift offers incredible opportunities to share information, build relationships, and engage with communities of shared interest. Specifically, the unique nature of cyberspace - which transcends geographic borders and communicative boundaries - has acted to democratise information production and consumption. Yet the features of this context also tend to foster polarising and negative behaviour, and this is exacerbated by the rise of artificial intelligence (AI), whereby algorithms amplify sensational, divisive content to maximise user engagement[1]. The possibility of being anonymous and not accountable for the harm caused by one's online behaviour has fueled widespread discriminatory trolling, threats, and harassment – all of which are disproportionately experienced by women and girls.

Research suggests that over half of all women and girls connected to the internet have personally experienced some form of online harassment, and this is even higher for women who are active in the public sphere, such as journalists, politicians, and human rights defenders [2,3,4]. In many of these attacks, women are discredited, dehumanised, and derogated utilising multimodal forms of gendered and sexualised harassment. For instance, creating and sharing pornographic deepfakes, undertaking campaigns of disinformation, and threatening physical and sexual violence, all of which are deliberately aimed at instilling fear, inflicting harm, and undermining gender equality efforts[5]. These types of targeted abuse function as a form of gendered censorship, systematically narrowing the space available for women's participation and inclusion online and causing widespread harm to victims[6].

These issues can be seen to belong to the domain of ***online misogyny***, which is broadly understood as expressions of hostility, contempt, or hatred toward women via digital technologies. The prevalence of this type of online violence has increased in parallel with the growth in digital connectivity, with recent research finding that 73% of young people report witnessing misogyny online, and 55% think it is increasing[7]. In fact, misogynistic content is often rewarded due to the effects of algorithmic amplification, whereby controversial and inflammatory content triggers high emotional arousal (both from supporters and critics) and because of high engagement, it is pushed to a wider audience[8,9]. The impacts of exposure to online misogyny can be severe, impacting the attitudes and behaviours of both men and women and making the online environment less safe for everyone. Also, digital hostility does not remain relegated to online contexts; it acts to legitimise offline gender-based discrimination and violence as well.

Despite the prevalence and impacts of online misogyny, efforts to counter these forms of violence at scale remain limited in effectiveness. A central but underappreciated reason for this is issues with the definition and operationalisation of this concept. Specifically, there is no agreed upon, cross-disciplinary understanding of online misogyny, what it encompasses, or how it should be measured for the purposes of detection or regulatory enforcement. Furthermore, the two research traditions that have engaged most directly with this problem, computer science and social science, have done so largely in isolation from one another, producing conceptual frameworks that are poorly aligned. Notably, although computational approaches have achieved technical sophistication in classifying overt forms of misogynistic content in online discourse, they tend to rely on narrow taxonomies that miss the subtler, multimodal, and contextually embedded forms of violence. In contrast, social scientific work has produced nuanced accounts of misogyny's functions, forms, and intersectional dimensions, but has not translated these into scalable detection tools or operationalisable policy frameworks.

The mismatch in approaches to online misogyny has practical consequences in terms of combating online violence. Specifically, regulations cannot be enforced, nor can regulators evaluate compliance against a set of standards where definitions remain contested. Therefore, addressing the problem of definition and operationalisation of online misogyny is critical for effective governance. To address this issue, the current research aims to explore distinct approaches to understanding and defining online misogyny, its manifestations, and how it perpetuates harm. Further, the brief highlights that inconsistencies in approaches to online misogyny affect how it is measured, detected, and prevented.

Methods

A scoping review of the published literature between 2015 - 2025 was conducted in databases such as Web of Science, Scopus, and Google Scholar using combinations of keywords (e.g., “online misogyny”, “online sexism”, “cyber-sexism”) across disciplinary fields including social psychology, gender studies, sociology, digital humanities, and digital communication. It is noted that this is not an exhaustive list of all the published research, but rather a selection across the main disciplines for the purpose of comparison. In total, 21 papers were identified and purposefully selected by the first author for inclusion. For each paper, the key information was extracted, including the definition of online misogyny, the discipline, the type of study, and the methodological approach. Additionally, the proposed solutions for detecting and reducing online misogyny were examined (if there were any) and synthesised across the literature. Full details are included in Table 1.

Findings/ Results of the Review

Definitions of Online Misogyny

Definitions of online misogyny across the social and computer sciences share a set of core attributes - hostility, contempt, or hatred directed at women through digital means - but diverge in terms of scope and emphasis. Computational definitions are designed for more concrete operationalisation. Therefore, they tend to define online misogyny in terms of its detectable features (i.e., slurs, insults, threats, and other explicit markers) and to organise these features into taxonomies suitable for annotation and classification. Anzovino and colleagues[10] developed a highly cited taxonomy that categorises misogynistic content into five types of discourse: discrediting, stereotyping and objectification, sexual harassment and threats of violence, dominance, and derailing. Guest and colleagues[11] developed a similar framework, defining misogyny as '*directed abuse at women or a closely related gendered group*.' They highlight four types of discourse: pejoratives (explicit, derogatory language), treatment (support for harm or disrespectful actions against women), derogation (implication of arguments of women’s inferiority to men), and personal attacks (weaponising gender traits). These taxonomies allow for the development of annotated datasets and the training of classification models to detect misogynistic content. However, they anchor misogyny in explicit, identifiable features, meaning they can miss the subtler, more implicit forms of misogynistic content that are increasingly prevalent in online spaces and that social scientists have documented extensively.

Social scientific definitions, in contrast, tend to be less functional and more inclusive; they define online misogyny as an extension of offline gendered violence against women, often taking the form of hostility, prejudice, and insults, aiming to denigrate, dehumanise, or discriminate against women. These definitions highlight that misogyny serves an identifiable purpose as part of a social system, particularly the regulation and constraint of women's participation in public life and the upholding of traditional gender norms, both of which have the outcome of reducing women’s voice and visibility in civic space[12]. In the book The Logic of Misogyny[13], by Kate Manne, misogyny is considered to have a policing mechanism to enforce sexism, which itself is the *justificatory* component of patriarchal ideologies. Online misogyny, therefore, seeks to utilise cyberspace to punish, discredit, and silence women who step outside prescribed gender roles (whether this be online or offline). A further component of social science definitions is the inclusion of intersectionality, where it is understood that online misogyny does not target all women equally. Women from diverse minority groups (e.g., cultural, sexual, disability, and religious) face distinct and

cumulative forms of gendered hostility. Computational models have been slow to incorporate intersectional perspectives, and existing detection systems have been shown to under-detect attacks against women of colour[14].

Disciplinary Approaches to Online Misogyny

Computer Science

In this discipline, online misogyny is treated through the lens of automated text classification, often reducing the task of identifying what it is to a binary decision: whether a given piece of text or discourse is misogynistic or not[17]. Following this approach, the concept is operationalised as a set of concrete characteristics that allow for classification. Computer science research on online misogyny detection has grown substantially over the past decade, driven by the availability of annotated datasets, advances in natural language processing (NLP: the field of computer science concerned with teaching machines to understand human language), and growing pressure to automate content moderation at scale.

This body of work has moved through generations: word and rule-based methods, machine learning models using engineered features, deep learning architectures using embeddings and, most recently, large language models (LLMs). Early systems relied on lists of misogynistic terms (lexicons) to identify behaviours, which worked reasonably well for overt abuse but struggled to generalise across platforms, languages, cultures, and styles of expression. The next generation integrated multiple feature types, including frequency weighting, sentiment scores, topic-specific vocabularies, and stylistic markers within more sophisticated statistical models. Research demonstrated that combining these features substantially outperformed any single approach, though performance remained sensitive to the type and source of training data[15,16]. Then large pretrained language models were used, including BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa, which acts as foundational language model designed to understand the deep context of words by looking at the sentences that surround them. This allows the system to learn richer representations of context and meaning and thus identify misogyny with greater nuance.

Despite this technical progress, there are still critical gaps. First, models trained on data from one platform do not generalise well to others: content norms, linguistic registers, and misogynistic vocabulary vary substantially across platforms such as X, Reddit, Facebook, and Instagram, and no single dataset or taxonomy has achieved cross-platform reliability. Second, detection is consistently stronger for overt and explicit forms of misogyny (i.e., slurs, explicit threats, graphic sexual content) than for the subtler forms that are often more prevalent in practice (i.e., sarcasm, irony, coded language)[17]. Third, existing systems perform markedly worse on non-English and non-Western instances of misogyny, which is a critical limitation given that platforms operate globally. Finally, intersectional forms of misogyny where women are targeted on the basis of gender combined with race, religion, sexuality, or disability, are under-detected by models.

Across all generations of research, there is no consensus on which type of model most effectively detects misogyny. Rather than the choice of algorithm, performance appears to depend most on representational richness, meaning that the use of contextual embeddings, diverse feature types, and varied training data consistently outperforms approaches that rely on text alone because they provide nuance[18]. Notably, this brief highlights that within computer science, online misogyny is primarily conceptualised as text-based classification problems solvable through optimised combinations of linguistic features, embeddings, and machine learning or deep learning architectures, rather than as complex socio-cultural constructs requiring broader contextual interpretation. What this also points to is the adoption of "simplistic" definitions of misogyny, as the task of capturing nuance or subtlety is difficult. Therefore, while computational approaches have produced a substantial body of research aimed at automating detection, this work remains largely disconnected from parallel developments in the social sciences.

Social Science

Social science has produced a rich and theoretically sophisticated body of knowledge about what online misogyny is, how it functions, whom it targets, and what its individual and social consequences are. This literature has established that online misogyny functions as part of a system of social control rather than a collection of discrete acts; that its harms are cumulative and structural, building through repeated exposure and sustained threats, rather than through any single incident; that it operates across modalities (text, image, video, audio), and that it is intersectional[23]. As described above, in this discipline, misogyny is seen as instrumental to the upholding the patriarchy via the moral and social policing of women and girls[12]. In the social sciences, this has also been conceptualised as *E-bile*[19], or language used in online environments that conveys hostility, trolls women, and bullies them for perceived violations of social norms. Furthermore, online misogyny is largely seen as coinciding with forms of gender-based harm (violence against women, hate speech, and misinformation about women and girls)[10,11,12].

This research has also produced the most substantive accounts of online misogyny's consequences. Women who experience sustained misogynistic abuse have reported significant psychological harm (e.g. anxiety, hypervigilance, shame, and post-traumatic symptoms) alongside concrete professional and civic costs (e.g. cyber-stalking, self-censorship, reduced online visibility, withdrawal from platforms)[19]. Importantly, misogyny is often situated within the wider frameworks of "violence against women", and "hate speech", whereby it is associated with harms such as making women targets of cyber/offline stalking, harassment, bullying, and threats of being physically attacked[20]. The research shows that the harms caused by online misogyny tend to be absorbed by women themselves while leaving the structural conditions that produce misogyny unaffected.

The key problem with social scientific approaches is that while they are conceptually rich, this does not readily translate into the kinds of scalable methods required to inform detection and regulatory enforcement. Specifically, definitions grounded in function, context, and impacts are powerful for explaining lived experience, but practically difficult to implement in detection and content moderation systems that must make rapid, automated decisions about a huge number of individual pieces of content to essentially perform (a binary) classification task to "detect" and "eliminate" misogynistic speech. This is not an unsolvable limitation of social scientific knowledge; it is a translation problem that has received insufficient attention in the literature. With deliberate effort, these concepts can be operationalised, and indeed some computer science taxonomies are already explicitly informed by feminist theory. The challenge is to do this systematically and collaboratively, rather than having computational researchers develop their own ad hoc operationalisations of social scientific concepts without the involvement of those who developed them.

Detecting and Reducing Online Misogyny

Across the literature reviewed, two types of actions to combat online misogyny were highlighted: (i) "DIY" Approaches, and (ii) systemic interventions. The first is an individual "DIY" approach, exemplified by Jane[20], who identifies this as the most commonly adopted, and paradoxically the most "effective" strategy, as it often results in women modifying their own online behaviour to reduce their visibility and perceived vulnerability to misogynistic attacks. DIY approaches include behaviours such as self-censorship to limit participation in public discourse, curating one's audience, or withdrawing from certain online spaces entirely. In addition, women frequently take measures to safeguard their offline identities to prevent doxxing, stalking, or other forms of targeted harassment, including using pseudonyms, frequently changing personal contact information, and altering residential details to stay safe[21]. While these strategies may offer short-term protection, they disproportionately burden women and contribute to the silencing of their voices in civic and digital spheres.

The second is systemic and structural interventions such as specialised training for law-enforcement agencies to recognise online misogyny as a form of hate-based conduct and the provision of legal protections for victims of online misogyny comparable to those applied to other forms of hate speech and misinformation campaigns[22]. Such perspectives highlight the need for robust regulatory and governance frameworks, institutional accountability, and social policy reform to meaningfully counter gendered digital violence. Yet it is critical that policy interventions must account for the unique features of online misogyny (as distinct from offline misogyny) by accounting for the potential anonymity of perpetrators, algorithmic amplification of content, and the cross-platform scale and reach. Computer science has attempted to develop tools/ datasets that address these features, which can be used to train AI models to detect (and subsequently address) online misogyny. Yet these tools need to be implemented in line with systematic approaches to technology governance.

The Policy Gap

The definitional and disciplinary gaps documented in this brief have critical implications for regulatory frameworks and governance. For example, some nations have instituted law that requires very large online platforms to identify systemic risks and to implement mitigation measures, explicitly including "cyber violence against women". Notably, the EU Directive on combating violence against women (2024), to be made into law by June 2027, criminalises some forms of online gender-based violence and sets minimum standards for legal responses. These regulations (and others across different countries) require that the types of violence they are supposed to address are clearly identified, but in the absence of shared definitional standards, compliance may be difficult, and enforcement inconsistent.

Current practice reflects this gap as reporting mechanisms on major platforms typically require users to categorise their experiences into predetermined options that do not reflect the multifaceted nature of online misogyny. Content moderators frequently lack gender-sensitive training and context, leading to inconsistent decisions and leaving women who report misogynistic abuse feeling that their experiences have been dismissed[23]. Furthermore, content moderation is also a global problem, and in contexts of low-resource languages, it remains severely underdeveloped, meaning that women from the Global South receive substantially less protection from platform moderation systems.

It is an urgent priority to co-develop a shared, interdisciplinary definitional framework for online misogyny that is both conceptually grounded and operationally useful. This should be structured so that core elements (the nature of the content, its target, its function, and its context) can be combined and applied flexibly across different platforms, regulatory contexts, and research purposes. It also needs to be multimodal, encompassing diverse types of content, rather than treating misogyny as a purely linguistic phenomenon. Intersectionality and nuance should be included to account for implicit, contextual, and cumulative forms of misogyny, and should be developed with dedicated resources for low-resource language annotation.

Conclusions

The results of this brief highlight that there are major disciplinary differences in definitions of and approaches to online misogyny. Computer science is concentrated around operationalising and specific misogynistic content and constructs, reflecting an orientation toward detecting discrete, measurable forms of misogyny online. In contrast, social scientific research adopts a more expansive, systemic, and intersectional perspective, engaging with themes such as gender roles, gender equality, identity, sexual violence, and the harms resulting from misogyny. This divergence results in a fragmented research landscape, where computational models risk oversimplifying the complexities of misogyny due to narrowly defined conceptual categories and social science risks not being able to effectively inform detection and mitigation measures. Consequently, improved integration between computational and social scientific approaches is essential. Such collaboration, focusing on the applicability and scalability of solutions, would promote more robust conceptual frameworks, enhance the ecological validity of detection systems, and ensure that automated tools reflect the nuanced, socio-cultural dimensions of gendered hostility rather than reducing it to isolated textual features.

Table 1: Literature included in the scoping review

	Citation	Discipline	Narrative Description of Online Misogyny	Type of Study	Method
1.	Alichie, B. O. (2023). "You don't talk like a woman": the influence of gender identity in the constructions of online misogyny. *Feminist Media Studies*, 23(4), 1409-1428. https://doi.org/10.1080/14680777.2022.2032253	Social Sciences	A form of violence against women in cyberspace that limits women's voices and visibility online, with feminist identity on social media acting as a key trigger for misogynistic targeting.	Qualitative	Thematic Analysis
2.	Al-Zaman, Md. S. (2021). Online Misogyny in Bangladesh. Asian Women, 37(3), 1–24. https://doi.org/10.14431/aw.2021.9.37.3.1	Social Sciences	Aggressive male reaction to women's presence in digital spaces, sustained through religious and cultural justifications that frame women's online participation as a moral transgression.	Qualitative	Content Analysis
3.	Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In: Silberztein, M., Atigui, F., & Kornyshova, E. et al. (Eds) *Natural language processing and information systems (NLDB) 2018. Lecture notes in computer science*, vol 10859. Springer, Cham, pp. 57–64	Computer Science	A specific case of hate speech targeting women in digital spaces, categorised into five types: discredit, stereotype and objectification, sexual harassment and threats of violence, dominance, and derailing.	Quantitative	Corpus annotation and automated classification
4.	Banet-Weiser, S. (2021). Misogyny and the politics of misinformation. In *The Routledge companion to media disinformation and populism* (pp. 211-220). Routledge.	Social Sciences	Anti-female violent expression circulating on popular media platforms, whose algorithmic logics and affordances amplify women's subordination and uphold male dominance online.	Commentary	Narrative
5.	Barker, K., & Jurasz, O. (2019). Online misogyny: A challenge for digital feminism? *Journal of International Affairs, 72*(2), 95–114. https://www.jstor.org/stable/26760834	Social Sciences	Gender-based abuse in digital spaces that uses public platforms to silence women and undermine gender equality, functioning as a tool of coercive control over women's online participation.	Commentary	Narrative
6.	Dafaure, M. (2022). Memes, trolls and the manosphere: mapping the manifold expressions of antifeminism and misogyny online. *European Journal of English Studies, 26*(2), 236–254. https://doi.org/10.1080/13825577.2022.2091299	Social Sciences	Historically rooted antifeminist discourse expressed through digital affordances (memes, irony, and trolling) and strategically reframed across platforms as a response to a perceived "crisis of masculinity."	Qualitative	Discourse Analysis

7.	Dickel, V., & Evolvi, G. (2023). "Victims of feminism": exploring networked misogyny and #MeToo in the manosphere. *Feminist Media Studies, 23*(4), 1392–1408. https://doi.org/10.1080/14680777.2022.2029925	Social Sciences	Networked misogynist narratives circulated across the manosphere (websites and social media groups united by the belief that men are oppressed by feminism) entangled with racist, homophobic, and far-right ideologies.	Qualitative	Thematic Analysis
8.	Dutta, A., Banducci, S., & Camargo, C. Q. (2025). Divided by discipline? A systematic literature review on the quantification of online sexism and misogyny using a semi-automated approach. *Scientometrics, 130*(9), 4915–4971. https://doi.org/10.1007/s11192-025-05410-2	Interdisciplinary	The enforcement of traditional patriarchal values in digital contexts, understood as a systematic political mechanism to dominate and control women, representing an extreme end of the sexism spectrum.	Review	Systematic Literature Review
9.	Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in Reddit. In Proceedings of the 10th ACM Conference on Web Science (pp. 87–96). *Association for Computing Machinery.* https://dl.acm.org/doi/10.1145/3292522.3326045	Computer Science	Hatred or contempt for women described as the "police force of sexism" operationalised as extreme and violent language across nine rhetorical categories spreading within and across manosphere communities.	Quantitative	Lexicon-based automated text analysis
10.	Fontanella, L., Chulvi, B., Ignazzi, E., Sarra, A., & Tontodimamma, A. (2024). How do we study misogyny in the digital age? A systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications, 11*(1), 1–15. https://doi.org/10.1057/s41599-024-02978-7	Social Science	Linguistically expressed hatred or aggression towards women in digital spaces, ranging from subtle exclusion and discrimination to severe sexual objectification and violent threats.	Review	Systematic Literature Review
11.	Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies*, 18(4), 515–524. https://doi.org/10.1080/14680777.2018.1447345	Social Sciences	Violence or harm directed at women online directly causing psychological, professional, or reputational harm, or indirectly making the internet a less equal, safe, or inclusive space for women.	Editorial	Narrative
12.	Guest E, Vidgen B, Mittos A et al (2021). An expert annotated dataset for the detection of online misogyny. In: *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*. Association for Computational Linguistics, pp. 1336–1350	Computer Science	A social problem making digital platforms toxic and unwelcoming to women, producing a silencing effect through distinct forms of abuse including gendered personal attacks, misogynistic pejoratives, and derogatory or threatening language.	Quantitative	Annotated dataset creation with automatic classification

13.	Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum, 30*(3), 284–297. https://doi.org/10.1080/10304312.2016.1166560	Social Sciences	Gendered "e-bile", a term coined in Jane's 2014 work for hostile gendered discourse online encompassing cyberbullying, cyber-stalking, and trolling. This paper argues that individual "digilante" responses to such abuse shift responsibility from perpetrators to targets and are insufficient as solutions.	Qualitative	Case study analysis
14.	Han, X. (2018). Searching for an online space for feminism? The Chinese feminist group Gender Watch Women's Voice and its changing approaches to online misogyny. *Feminist Media Studies, 18*(4), 734–749. https://doi.org/10.1080/14680777.2018.1447430	Social Sciences	Hostile messaging confronting feminist activists in digital spaces, shaped by patriarchal ideology and platform affordances, and, in the Chinese context, intensified further by state censorship and surveillance of online feminist activity.	Qualitative	Textual and content analysis
15.	Manne, Kate, (2017). *Down Girl: The Logic of Misogyny*. Oxford Academic.	Social Sciences	The enforcement branch of patriarchy that polices women's behaviour online and offline, punishing those who violate patriarchal norms and rewarding compliance and functioning alongside sexism to regulate women's conduct rather than simply expressing hatred.	Various	
16.	Mantilla, K. (2015). *Gendertrolling: How Misogyny Went Viral. Santa Barbara: Praeger*.	Social Sciences	"Gendertrolling" a coordinated, virally amplified form of digital harassment that exploits networked affordances to silence women's public voices and drive them from online civic discourse through mass-scale intimidation.	Various	
17.	Mohasseb, A., Amer, E., Chiroma, F., Tranchese, A. (2025) Leveraging Advanced NLP Techniques and Data Augmentation to Enhance Online Misogyny Detection. Appl. Sci.,15, 856. https://doi.org/10.3390/app15020856	Computer Science	The use of internet technologies to target and harm women through hostile and sexist language, whose dynamic and context-dependent nature makes it difficult for traditional digital moderation methods to address.	Quantitative	Transformer-based text classification with data augmentation

18.	Moloney, M.E, Love, T.P. (2018). Assessing online misogyny: Perspectives from sociology and feminist media studies. *Sociology Compass*.,12:e12577. https://doi.org/10.1111/soc4.12577	Social Sciences	Hatred or contempt for women expressed in digital spaces, reviewed through key conceptualisations including online sexual harassment, gendertrolling, e-bile, and disciplinary rhetoric.	Editorial	Narrative
19.	Parikh, P., Abburi, H., Chhaya, N., Gupta, M., & Varma, V. (2021). Categorizing Sexism and Misogyny through Neural Approaches. *ACM Transactions on the Web, 15*(4), 1–31. https://doi.org/10.1145/3457189	Computer Science	Entrenched prejudice or hatred directed at women, operationalised computationally as language actively perpetrating hostility, objectification, or dominance in digital spaces — treated as distinct from, but related to, sexism.	Quantitative	Neural text classification
20.	Srivastava, K., Chaudhury, S., Bhat, P., & Sahu, S. (2017). Misogyny, feminism, and sexual harassment. Industrial Psychiatry Journal, 26, 111.	Social Sciences	Hatred or contempt for women arising from patriarchal systems, expressed online and offline through sexual harassment and other mechanisms that enforce gender-based power imbalances.	Editorial	Narrative
21.	Philine Zeinert, Nanna Inie, and Leon Derczynski. (2021). Annotating Online Misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.	Computer Science	A category of abusive language in digital spaces with serious social consequences, the automatic detection of which is challenged by its linguistic complexity and diversity across platforms.	Quantitative	Expert annotation and taxonomy development

REFERENCE

1. Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2023). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus,* Volume 4, Issue 3, March 2025, pgaf062, https://doi.org/10.1093/pnasnexus/pgaf062
2. Plan International. (2020). Free to Be Online? Plan International. https://plan-international.org/publications/free-to-be-online
3. UNESCO. (2020). UNESCO's Global Survey on Online Violence against Women Journalists. Unesco.org. https://www.unesco.org/en/articles/unescos-global-survey-online-violence-against-women-journalists
4. Inter-Parliamentary Union. (2016). Sexism, harassment and violence against women parliamentarians (Issue Brief No. 2016-10). IPU. https://www.ipu.org/resources/publications/issue-briefs/2016-10/sexism-harassment-and-violence-against-women-parliamentarians
5. Baekgaard, K. (2024). Technology-facilitated gender-based violence. Georgetown Institute for Women, Peace and Security / Prevention Collaborative. https://prevention-collaborative.org/wp-content/uploads/2024/12/Technology-Facilitated-Gender-Based-Violence.pdf
6. Amnesty International. (2018). Troll Patrol findings: Using crowdsourcing, data science & machine learning to measure violence and abuse against women on Twitter. Amnesty International. https://decoders.amnesty.org/projects/troll-patrol/findings
7. Amnesty International UK. (2025). Toxic tech: New polling exposes widespread online misogyny driving Gen Z away from social media. https://www.amnesty.org.uk/documents/104/Amnesty20International20-20Gen20Z20and20Online20Misogyny20202420-20Savant_15tE7g9.pdf
8. Regehr, K., Kent, C., & Ringrose, J. (2024). Safer scrolling: How social media algorithms amplify misogynistic content to school-aged boys. UCL Centre for Sociology of Education / University of Kent. https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf
9. Over, H., Bunce, C., Konu, D., & Zendle, D. (2025). Editorial Perspective: What do we need to know about the manosphere and young people's mental health? *Child and Adolescent Mental Health, 30*(3), 272–274. https://doi.org/10.1111/camh.12747
10. Anzovino M, Fersini E, Rosso P (2018) Automatic identification and classification of misogynistic language on Twitter. In: Silberztein M, Atigui F, Kornyshova E et al (eds) *Natural language processing and information systems (NLDB) 2018. Lecture Notes in Computer Science, 10859*. Springer, Cham, pp. 57–64
11. Guest E, Vidgen B, Mittos A et al An expert annotated dataset for the detection of online misogyny. In: Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1336–1350 (2021)
12. Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies, 18*(4), 515–524. https://doi.org/10.1080/14680777.2018.1447345
13. Manne, Kate, (2017). *Down Girl: The Logic of Misogyny.* Oxford Academic. https://doi.org/10.1093/oso/9780190604981.001.0001
14. Thakur, D. (2024, October 2). Hated more: Online violence targeting women of color candidates in the 2024 US election [Report]. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2024/10/2024-10-02-CDT-Research-Hated-More-brief.pdf
15. Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems, 36*(5), 4743–4752. https://doi.org/10.3233/JIFS-179023
16. Attanasio, G., & Pastor, E. (2020). PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identification in Italian tweets. In V. Basile, D. Croce, M. Maro, & L. C. Passaro (Eds.), EVALITA Evaluation of NLP and Speech Tools for Italian – December 17th, 2020. Accademia University Press. https://doi.org/10.4000/books.aaccademia.6807
17. Sheppard, B., Richter, A., Cohen, A., Smith, E. A., Kneese, T., Pelletier, C., Baldini, I., & Dong, Y. (2024). Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 427–452). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.24

18. Dutta, A., Banducci, S. & Camargo, C.Q. (2025). Divided by discipline? A systematic literature review on the quantification of online sexism and misogyny using a semi-automated approach. *Scientometrics 130,* 4915–4971. https://doi.org/10.1007/s11192-025-05410-2
19. Posetti, J., Aboulez, N., Bontcheva, K., Harrison, J., & Waisbord, S. (2020). Online violence against women journalists: A global snapshot of incidence and impacts [Report]. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000375136
20. Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum, 30*(3), 284–297. https://doi.org/10.1080/10304312.2016.1166560
21. Alichie, B. O. (2023). "You don't talk like a woman": The influence of gender identity in the constructions of online misogyny. *Feminist Media Studies, 23*(4), 1409–1428. https://doi.org/10.1080/14680777.2022.2032253
22. Mantilla, K. (2015). Gendertrolling: How Misogyny Went Viral. Santa Barbara: Praeger.
23. Griffin, R. (2022, November 1). An intersectional lens on online gender based violence and the DSA. Verfassungsblog. https://verfassungsblog.de/dsa-intersectional/

EDITORIAL INFORMATION

Author biographies
Isha Mangurkar is a social psychologist and Visiting Research Fellow at the United Nations University in Macau and tutor at the University of Edinburgh. Her research examines the intersection of gender-based harm, social identity, and online interaction, with a focus on how digital platforms enable and amplify misogyny across multiple countries and platforms.

Jaimee Stuart is Head of Research (Acting) at United Nations University Institute in Macau. Jaimee is an applied cultural and developmental psychologist who specialises in digital health and wellbeing.

Corresponding author
Jaimee Stuart (stuart@unu.edu)

Disclaimer
The views and opinions expressed in this paper do not necessarily reflect the official policy or position of the United Nations University.

Citation
Isha Mangurkar, Jaimee Stuart, Defining and Detecting Online Misogyny: Aligning Social and Computer Sciences, UNU Macau Research Brief 2 (Macau: United Nations University Institute in Macau, 2026).