BiasShield: An AI Browser Extension Against Online Misogyny

Filipa Chambel Vieira
Brunel University of London
London, United Kingdom
2223495@brunel.ac.uk

Cigdem Sengul
Brunel University of London
London, United Kingdom
cigdem.sengul@brunel.ac.uk

Abstract

Online spaces frequently expose women to sexualised and objectifying content, with documented harms including body dissatisfaction, anxiety, and depression. Automated moderation algorithms compound this through gendered bias by disproportionately classifying benign images of women as sexualised. Deepfake technologies have intensified the harms, with the victims being predominantly women. To counter these developments, we present BiasShield, a browser extension that identifies, audits, and enables users to manage exposure to misogynistic and deepfake content. We report on the design of a multimodal classifier and evaluate its capacity to detect misogynistic content while reducing gender-based false positives. By making algorithmic bias visible and actionable through exposure analytics and protective measures- including optional blurring of offensive content—BiasShield turns content moderation on the web into informed, user-based control.

CCS Concepts

• **Human-centered computing** → **Web-based interaction**; • **Social and professional topics** → **Gender**; • **Information systems** → **Personalization**; *Document filtering*.

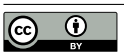
Keywords

Misogyny detection, Deepfakes, User empowerment

ACM Reference Format:
Filipa Chambel Vieira and Cigdem Sengul. 2026. BiasShield: An AI Browser Extension Against Online Misogyny. In *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3795513.3810444

1 Introduction

Women are systematically exposed to online misogyny via objectifying and sexualized content, with documented psychological and social harm [10]. Technology companies have developed algorithms to moderate such content, but these systems embed gendered assumptions that produce discriminatory outcomes. Mauro and Schellman (2023) demonstrated this directly: when Mauro (a white male) was photographed with a bare chest and then with a bra, Azure Vision assigned raciness scores of 22% and 97%, respectively [9]. Indeed, automated moderation systems end up enforcing misogyny, disproportionately penalising women's content, marking neutral images as sexual more often [12].



WebSci Companion '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2492-3/26/05
https://doi.org/10.1145/3795513.3810444

Deepfake technologies have intensified this harm by enabling the creation of non-consensual sexualized imagery of women, resulting in psychological trauma, reputational damage, and online harassment [8]. Deepfake pornography disproportionately targets women: reports indicate that 98% of all deepfake content was pornographic in 2023, with 99% of victims being women [1].

These issues produce a twofold harm: over-moderation, where benign content associated with women is censored, silencing self-expression and cultural representation; and under-protection, where the objectifying content remains visible.

To give users direct control over their web experience, this paper proposes BiasShield, a browser-based AI extension that identifies, audits, and manages exposure to misogynistic and deepfake content through analytics and protective features such as optional blurring of offensive text and images. Thus, BiasShield addresses both failure modes — reducing false positives on benign content while detecting and flagging genuinely harmful material — placing moderation decisions in the hands of the user.

2 Background and Related Work

Detecting misogyny online is challenging due to the scarcity of high-quality annotated data [4]. To address this, Guest et al. developed an expert-annotated Reddit dataset with a hierarchical taxonomy of misogyny categories, which comprise: pejoratives (terms demeaning women), misogynistic treatment (content inciting harm), derogation (content that belittles women), and gendered personal attacks. Their logistic model achieved high precision at the cost of low recall, while a class-weighted BERT provided a more balanced but still moderate F1 of 0.43 on misogynistic cases. The MISTRA framework improved on this (Macro-F1 71.5%) by fusing features across image, text, and generated caption modalities [6].

Text-only models often fail when evaluating decontextualised content. Hebert et al. [5] show that a holistic analysis of text and images grounded in the surrounding discussion context, using a Multi-Modal Discussion Transformer (mDT), outperforms several text-only baselines, including BERT-HateXplain and Detoxify[1].

Commercial moderation tools detect harmful content across text, images, audio, and video, but offer users limited transparency and agency — this is the gap that BiasShield aims to address.

3 BiasShield Design

BiasShield is implemented as a Chromium browser extension and an inference backend. A core design principle is that toxicity is often contextual: a text or an image that appears ambiguous may become abusive when interpreted along with its surrounding text, neighbour elements, or the broader discussion in which it appears [5]. Hence, BiasShield extracts website components, such as text

[1] https://huggingface.co/tum-nlp/bert-hateXplain; https://github.com/unitaryai/detoxify

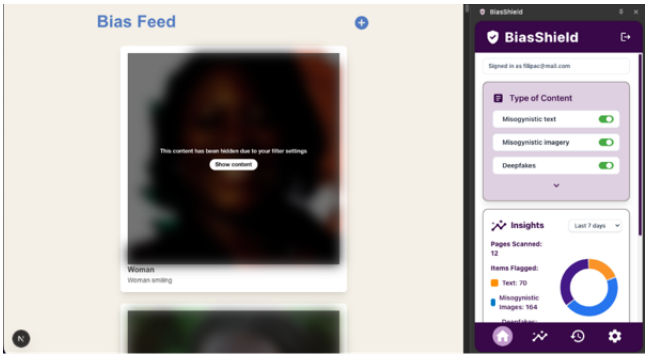


Figure 1: BiasShield in action: a detected deepfake image is blurred according to user preferences, while the text remains unaffected. The analytics panel displays misogyny exposure metrics for the current browsing session.

blocks and images, into context elements and sends compressed representations to the backend for analysis.

The backend returns predictions, which are then used to apply user-configurable actions on the page to filter content. The extension provides a dashboard with insights into flagged content, an activity log with information on the sites visited during the session, and a settings panel where the user can configure the active detectors and filter options, such as "blur content" and "reveal-on-click".

BiasShield combines a text classifier for detecting misogynistic discourse with an image classifier comprising two heads: misogyny detection and deepfake detection. The project builds on different datasets, including human images [11], GTS human fashion images [7], memes [3], deepfakes [2] and Reddit posts and comments [4]. The text classifier uses the DistilBERT model fine-tuned on Reddit data [4] with focal loss to address class imbalance. The image misogyny head was trained on a combined dataset of misogynistic memes and neutral human fashion images [3, 7, 11], decoupling the presence of women from misogynistic classification. A separate deepfake detection head [2] flags synthetic imagery.

4 Findings and Discussion

BiasShield reframes content moderation as a user-controlled process rather than a platform-enforced one. The system enables transparency by showing what is flagged and why, analytics, and controllable mitigation options (blur/reveal-on-click) (see Figure 1).

We evaluated BiasShield on BiasFeed, a website comprising text and images from the meme, neutral human-image, deepfake and Reddit test data. Examples of correctly detected misogynistic content include images of suited men with overlaid text dismissing women in the workplace, and images with text normalising domestic violence. In both cases the multimodal classifier correctly identified the misogynistic signal conveyed through the combination of image context and overlaid text. However, misclassification errors occur due to categorising unexplicit misogynistic texts as non-misogynistic. On the Reddit test set, the text classifier achieved an F1 of 0.65, a precision of 0.96, and a recall of 0.49. Challenging examples include: "The way every man feels when a woman is driving" or "Women drivers. I bet the boat just jumped out in front of you." These statements often rely on stereotypes or irony rather than explicit slurs, presenting a categorisation challenge.

The ablation study confirmed that misogyny detection is primarily a language-driven multi-modal task: the image-only variant achieved just 22.1% F1, while configurations incorporating overlaid text and generated context consistently outperformed it. The image classifier incorporated a Variational Autoencoder (VAE) to normalise and compress visual representations across heterogeneous web images. However, the study showed this compression discarded discriminative information: the configuration without the VAE achieved the best F1 of 94.9%.

Fairness evaluation used benign human images split into female-target and non-female-target groups. As all neutral images were non-misogynistic, any positive prediction counted as a false positive. Before image regularisation, 34.5% of female-targeted images were incorrectly flagged as misogynistic, compared with 1% of non-female-targeted images, a gap of 33.5%. A final fine-tuning stage on neutral images reduced false positives to 0% in both groups, with a small trade-off on the meme test set, where macro-F1 fell from 0.931 to 0.919 and misogynistic recall from 0.913 to 0.875.

Overall, the results indicate that overlaid text and generated context are complementary — neither alone is sufficient — and that the multimodal design is essential for capturing the linguistic and contextual signals through which misogyny is typically expressed, whether in images or text.

References

[1] [n. d.]. 2023 State of Deepfakes. https://www.securityhero.io/state-of-deepfakes/
[2] A. Gamal. 2025. OpenForensics: Wild Dataset. https://www.kaggle.com/datasets/ahmedgamal/openforensics-wild-dataset
[3] F. Gasparini, G. Rizzi, A. Saibene, and E. Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief* 44 (2022), 108526. doi:10.1016/j.dib.2022.108526
[4] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts. 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*, P. Merlo, J. Tiedemann, and R. Tsarfaty (Eds.). 1336–1350. doi:10.18653/v1/2021.eacl-main.114
[5] Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2024. Multi-Modal Discussion Transformer: Integrating Text, Images and Graph Transformers to Detect Hate Speech on Social Media. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 20 (Mar. 2024), 22096–22104. doi:10.1609/aaai.v38i20.30213
[6] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi. 2024. MISTRA: Misogyny Detection through Text–Image Fusion and Representation Analysis. *Natural Language Processing Journal* 7 (2024), 100073. doi:10.1016/j.nlp.2024.100073
[7] W. Khan. 2024. Fashion Gender Classification Dataset for Machine Learning. https://www.kaggle.com/datasets/engrmwaqasniazi/mens-and-womens-images-for-fashion-classification/data
[8] B. U. Mahmud and A. Sharmin. 2023. Deep Insights of Deepfake Technology: A Review. arXiv:2105.00192 [cs.LG] https://arxiv.org/abs/2105.00192
[9] G. Mauro and H. Schellman. 2023. 'There is no standard': investigation finds AI algorithms objectify women's bodies. https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies
[10] J. S. Hyde S. Grabe, L. M. Ward. 2008. The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychological Bulletin* 134, 3 (2008), 460–476. https://doi.org/10.1037/0033-2909.134.3.460
[11] M. Sanaei. 2024. Human Images Dataset - Men and Women. https://www.kaggle.com/datasets/snmahsa/human-images-dataset-men-and-women/data
[12] C. Skopelity. 2024. 'There are some really extreme views': young people face onslaught of misogyny online. https://www.theguardian.com/society/2024/mar/01/there-are-some-really-extreme-views-young-people-face-onslaught-of-misogyny-online