PREVENTING TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE IN EGYPT

AN AI-POWERED RESEARCH TOOL

KVINFO

KVINFO
KØN OG LIGESTILLING

LIST OF CONTENT

This publication is part of the acceleration project *Digital Rights: Towards a more inclusive online space*, by International Media Support (IMS) and KVINFO, in collaboration with Jordan Open Source Association (JOSA) and Tadwein in the HRIC consortium, in the DAPP programme. Written by KVINFO

FOREWORD

The digital sphere has become one of the most influential arenas for public debate, civic participation, and democratic engagement. Across the Middle East and Northern Africa (MENA) region, online platforms offer unprecedented opportunities for voices hitherto excluded such as women, youth, and communities made marginalised, to express opinions and demands, aims and aspirations, to mobilise and organise, as well as to create and influence public discourse, and participate in public and political debates and activism.

However, due to increasing gender-based hate speech, harassment, and other forms of technology facilitated gender-based violence (TFGBV)[1], this development, has highly ambiguous consequences for gender justice, including the right to equal participation, the right to freedom of expression[2] and a life free from threats and violence.

TFGBV affects millions of people, and the impact of TFGBV on women, girls, LGBTIQ+ persons, and other specifically exposed groups is profound and multifaceted. In a study from 2021, 60% of women in the Arab states reported experiencing some form of online harassment.[3] Globally, different studies indicate that between 16-60% of women are affected.[4] Looking specifically at women human rights defenders, activists, journalists and media workers, the number is as high as 70%.[5] Perpetrators are often associated with anti-gender and anti-feminist individuals or groups, but can also be state agents or other users of ICT.

This is why International Media Support (IMS), Jordan Open Source Association (JOSA), KVINFO and Tadwein set out to create an AI-powered tool for gendered hate speech detection in 2025, focusing on the Egyptian dialect of Arabic. The project has included collecting experiences from a pilot project in Jordan, a robust annotation process, focus group discussions with experts, training of trainers and many more elements, in order to produce an AI-model with high precision of detection and intuitive useability. The model includes features facilitating use for a range of tech savvy to non-tech interested activists, researchers, journalists and organisations.

This publication includes some of our experiences and reflections.

---

[1] In this project, and as organisations, we lean on UNDP's definition of TFGBV. Please see chapter 2 for the full definition.

[2] Special Rapporteur on the promotion and protection of freedom of opinion and expression (2021) *Gender justice and freedom of expression*. UN General Assembly, A/76/258 A/76/258: Gender justice and freedom of expression - Report of Special Rapporteur on the promotion and protection of freedom of opinion and expression | OHCHR

[3] UN Women, 2021, *Violence against women in the online space: insights from a multi-country study in the Arab States* https://arabstates.unwomen.org/en/digital-library/publications/2021/11/violence-against-women-in-the-online-space#view

[4] United Nations, Report of the General Secretary, 2024, *Intensification of efforts to eliminate all forms of violence against women and girls: technology-facilitated violence against women and girls* a-79-500-sg-report-ending-violence-against-women-and-girls-2024-en.pdf

[5] Posetti, J. et. al, 2025, *Tipping point: The chilling escalation of online violence against women in the public sphere*, UN Women, tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

1. INTRODUCTION

Online hate speech has become a global concern, amplified by the rapid expansion of social media platforms and the limited effectiveness of moderation systems, combined with a lack of willingness even to moderate properly, as many social media platforms' business models equals large amounts of traffic, which you tend to achieve by promoting polarized views. In the MENA region, online abuse often intersects with deeply rooted gender norms, political polarisation, and identity-based discrimination. Women who participate in public debate—journalists, activists, politicians, and young users—are particularly exposed to targeted harassment, threats, and misogynistic narratives.[6]

Gendered and sexist hate speech frequently overlaps with other forms of discrimination, including racism, xenophobia, and religious intolerance. This convergence exacerbates harm, discourages participation, and undermines fundamental human rights, including freedom of expression and access to information, and thereby democracy.

Civil society organisations, journalists and researchers play a critical role in documenting and reporting online harm, supporting affected individuals, and advocating for policy change. They need more leverage, not least in the shape of quantitative data, in order to catalyse systemic changes. The scarcity of updated data and comprehensive studies also speak to the need of further research, and not least quantitative studies. However, the sheer volume and velocity of online content make manual monitoring insufficient. AI-powered tools offer new possibilities to systematically detect patterns of abuse, generate evidence, activate protection mechanisms and strengthen advocacy efforts.

This publication aims to:

- Present an AI-powered research tool designed to detect online gendered hate speech in Egyptian Arabic;
- Assess the usability, effectiveness and limitations of AI tools for CSOs, journalists and researchers;
- Provide recommendations for civil society, researchers, policymakers, newsrooms and other stakeholders when using AI-powered research tools.

JOSA has developed a similar AI-model for Jordanian Arabic ("Nuha") in parallel to this project. During the training phase of Nuha, JOSA observed online gender-based violence (GBV) targeting women activists. Comments on social media platforms directed at women activists sought to intimidate and discourage them from participating in politics and activism. Misogynistic irony and sarcasm was employed to bypass automated content moderation. Cases documented by JOSA indicated that women politicians and activists, including feminist activists, were more subjected to hate speech online. Hate speech against women in politics in Jordan included demeaning comments, accusations of importing and implementing so called

[6] The International Center for Journalists' "Big Data Case Study" examines the prolific campaign of online violence against journalist Ghada Oueiss in the context of her lived experience of the abuse, her journalistic work, and the socio-cultural and political conditions in which the abuse occurs, ICFJ_BigData_Ghada Oueiss_Online Violence.pdf

"Western agendas", and the use of negative religious discourse and gender-based stereotypes. Not even women ministers were free of TFGBV.[7]

2. UNDERSTANDING ONLINE HATE SPEECH AND TFGBV

While definitions vary across academic and policy contexts, hate speech generally refers to expressions that promote, justify, or incite discrimination, hostility, or violence against individuals or groups based on protected characteristics such as gender, ethnicity, religion, nationality, or sexual orientation. In the context of this project, particular attention is given to gendered hate speech and other forms of TFGBV, which targets individuals—most often women and LGBT+ people—because of their actual or perceived gender.

KVINFO applies UNFPA's definition for gender-based violence in cyberspace:

> *"Technology-facilitated gender-based violence, or TFGBV, is an act of violence perpetrated by one or more individuals that is committed, assisted, aggravated and amplified in part or fully by the use of information and communication technology (ICT) or digital media, against a person on the basis of their gender."*[8]

This definition is used in order to emphasise how gender-based violence in cyberspace results in or is likely to result in physical, sexual, psychological, social, political, and/ or economic harm, or other infringements of rights and freedoms, leading to withdrawal from public and political participation, thus causing a democratic deficit. See also KVINFO's strategy on TFGBV, annexed to this publication.

MANIFESTATIONS

Online gendered hate speech manifests in multiple forms and degrees of severity. These range from sexist language, misogynistic stereotypes, and sexualised insults to threats of violence, calls for exclusion, coordinated disinformation and harassment campaigns. Such expressions, whether intentional or not, often lead to the silencing of certain groups and individuals, discourage their participation in public debate, and reinforce existing power imbalances. Importantly, gendered hate speech does not exist in isolation. In many instances, it intersects with other identity-based forms of abuse, including racism, xenophobia, religious intolerance, or political persecution. This intersectionality intensifies harm and disproportionately affects women and other persons who belong to multiple marginalised groups.

---

[7] Democracy Reporting International (2024) *Online Public Discourse in the MENA Region, Persistent Tactics of Disinformation and an Increase in Online Gender-Based Violence*, 65b21f4539b6e.pdf

[8] Coined by UNFPA, see https://www.unfpa.org/TFGBV, see also UN Women *Expert Group Meeting report: Technology-facilitated violence against women: Towards a common definition*, https://www.unwomen.org/en/digital-library/publications/2023/03/expert-group-meeting-report-technology-facilitated-violence-against-women

Across digital platforms, observations indicate that gendered hate speech is both widespread and context-dependent. In the MENA region, and particularly in Egypt, online attacks against women tend to intensify around politically sensitive moments, such as elections, legislative reforms, or high-profile debates on women's rights. Social media platforms—including Facebook, X (formerly Twitter), TikTok, and Instagram—serve as primary spaces where such abuse unfolds, due to their reach, speed, and algorithmic amplification of polarising content. Public figures, activists, journalists, and young women engaging in social or political discussions are especially exposed to targeted attacks.[9]

A CASE FROM DENMARK

Also, beyond the MENA region, public debate increasingly occurs online. In Denmark, where KVINFO is based, almost four times as many citizens participate in public debate online compared to public debates in physical spaces.[10] This development places huge responsibility on platform moderation, especially of pages belonging to political parties, politicians and media outlets. According to a report by the Danish Institute for Human Rights, far more women than men feel the so-called chilling effect of potential TFGBV and online harassment, and refrain from participating in public online debates.[11] The fear is not unfounded, seeing as far more women than men are exposed to hateful attacks and comments. This is illustrated not least by a unique Danish study, based on 73 million posts and comments on pages belonging to Danish media and politicians on Facebook over a period of just above three years. According to the study, 73 percent of gender-based hate was directed towards women, 27 percent against men and five percent against gender minorities. Additionally, and importantly, an estimated 68 percent of perpetrators in public debate were written by male-named profiles and only 32 percent by female-named profiles.[12] This speaks to patriarchal structures being reproduced online. The chilling effect is also documented in Jordan, not least in a study by the Renaissance Strategic Centre, reporting that 77.4% of women who experience online violence responded by self-censoring, anonymizing their profiles, or withdrawing from platforms entirely to avoid further attacks.[13]

Another troubling aspect of sexism in digital spaces is the more hidden, so called "everyday sexism", a topic that KVINFO has worked with quite extensively in a Danish context. Everyday sexism is more subtle than threats or attacks. Examples include comments that build on stereotypical ideas of gender, patronising remarks or misogyny disguised as jokes. On social media, these include likes, shares and emojis that support sexist messages. This may seem

[9] See for example Online gender-based violence: What scenario for the MENA region? - EuroMed Rights
[10] Zuleta, L. et al, *Ytringsfrihed og selvcensur*, DIHR, 2024, Ytringsfrihed-og-selvcensur_DK_juni2024.pdf
[11] Danish Institute for Human Rights (DIHR) *DEBAT: Demokratiet lider, når kvinder flygter fra den offentlige debat*, DEBAT: Demokratiet lider, når kvinder flygter fra den offentlige debat | Institut for Menneskerettigheder
[12] Analyse & Tal, KVINFO and TrygFonden (2021) *Angreb & had i den offentlige debat på Facebook,* Angreb-og-had-i-den-offentlige-debat-paa-Facebook.pdf
[13] Renaissance Strategic Centre (2026) *How TFGBV Impacts the Democratic Participation of Young Women: Jordan as a Case Study,* TFGBV-report.pdf

harmless, but are part of creating a culture of sexism that allows TFGBV to flourish. Everyday sexism can contribute to the feeling of not belonging in public spaces and debates.[14]

ONLINE OFFLINE CONTINUUM

The urgency of addressing sexist hate speech and tackling online gendered disinformation is a topic prioritized by IMS. Online violence has real life consequences, similar to those of violence in the physical world. For example, in Pakistan, nine out of 10 women journalists reported that their mental health has been affected, and globally cases of stress, anxiety, depression, post-traumatic stress and suicide have been recorded. As a result, women journalists, who are already underrepresented in the news, use various forms of self-protection strategies such as turning down assignments, working under pseudonyms, not sharing opinions online, changing their beat to different topics or leaving the profession. Online gender-based violence is not just a private violation but a real and acute threat to freedom of expression.[15]

Research emphasises that TFGBV and offline violence exist on a continuum rather than as separate phenomena. Digital abuse can serve as a precursor to physical violence: online threats, doxing, and sharing of intimate images have been linked to stalking, physical assault, sexual coercion, and other forms of offline harm.[16] Empirical evidence from KVINFO's work in Egypt sustains this finding, and, for example, points at the use of TFGBV to hinder divorce. Research on intimate partner violence shows that TFGBV commonly co-occurs with offline coercive and controlling behaviours, reinforcing the role of digital tactics in broader patterns of violence and abuse.[17]

Although systematic data on magnitude and causal pathways remain limited, the growing body of evidence underscores that digital violence is not "just online"—it can escalate into real-world harm and reflects and reinforces offline gender-based violence patterns.[18] According to new research from UN Women, among 640 women human rights defenders, activists, journalists and media workers surveyed, 42% had experienced offline harm triggered by online violence, which is more than double the incidence rate recorded in 2020.[19]

Another aggravating factor is that the vast majority of persons exposed to TFGBV do not report it to the competent authorities (such as police and Anti-Cyber Crimes Units) due to lack

---

[14] Analyse & Tal, KVINFO and TrygFonden (2025) *Digital hverdagssexisme: Hvordan forebygger og håndterer man den hårde tone i onlinedebatter,* https://kvinfo.dk/wp-content/uploads/2025/04/Digital-hverdagssexisme_20250430.pdf

[15] International Media Support (IMS) *Online gendered disinformation and sexist hate speech*, IMS-Online-gendered-disinformation_final.pdf

[16] Jordan Open Source Association (JOSA) and KVINFO, *The Impact of Technology-Facilitated Gender-Based Violence (TFGBV) on Survivors in Jordan*, Final-JOSA-KVINFO-TFGBV-Study-1.pdf

[17] Storey, J. E. et al. (2021) 'Technology-Facilitated Intimate Partner Violence: An Examination of Prevalence, Perpetration Type and Methods and the Impact of COVID-19', *Journal of Interpersonal Violence*, Technology-Facilitated Intimate Partner Violence: An Examination of Prevalence, Perpetration Type and Methods and the Impact of COVID-19 - PubMed

[18] UN Women and UNDP, *Technology Facilitated Gender-Based Violence: Developing a shared research agenda*, technology-facilitated-gender-based-violence-shared-research-agenda-en.pdf

[19] Posetti, J. et. al, 2025, *Tipping point: The chilling escalation of online violence against women in the public sphere*, UN Women, tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

of awareness, social stigmatisation, bureaucratic procedures, overall lack of trust in public institutions in handling sensitive cases and lack of legal frameworks and precedence, etc. Some forms of TFGBV have led to honour-related femicides, where women are killed by a family member/s.[20] This includes women who seek support from family members.[21]

SELF CENSORHIP AND ONLINE MODERATION

The characteristics of TFGBV present substantial challenges for those seeking to identify, monitor, and respond to hate speech. Automated systems struggle to interpret context, intent, and nuance, while human moderation alone would be overwhelmed by the scale and volume of online content. As a result of the current climate, many users, including journalists[22], self-censor or even withdraw from online spaces. In Jordan, a survey found that 77.4% of women who experience online violence "respond by self-censoring, anonymizing their profiles, or withdrawing from platforms entirely to avoid further attacks."[23]

Research shows that online moderators are key to combating gendered hate speech and discrimination in online debates. At the same time, they struggle to identify and handle expressions of everyday sexism. Hateful comments and threats are more easily distinguishable, and with violent forms of discrimination being part of their working day, there is a normalisation of everyday sexism. This kind of normalisation, that happens also for online users, lays the foundation for aggravated violence. As one moderator in the study puts it:

> *"Sexism has become so ingrained in our comments that we no longer notice it. We don't spot it immediately when we have to read a lot of comments. Racist expressions, on the other hand, seem more violent, and we also react to them more quickly."*[24]

One of the defining characteristics of online hate speech is its adaptability. Language used to convey hate evolves rapidly, often relying on coded expressions, humour, irony and sarcasm, or culturally specific references that make detection difficult. In Arabic-speaking contexts, this challenge is further compounded by the diversity of dialects, transliterations, and the mixing of Arabic with other languages such as English or French. What may appear neutral at a surface level can, within a specific cultural or political context, carry deeply offensive or threatening connotations.

Understanding hate speech online and TFGBV therefore requires not only definitions but also a contextual and human rights–based perspective. It demands attention to how power,

---

[20] Democracy Reporting International (2024) *Online Public Discourse in the MENA Region, Persistent Tactics of Disinformation and an Increase in Online Gender-Based Violence*, 65b21f4539b6e.pdf p.65 and Jordan Open Source Association (JOSA) and KVINFO, *The Impact of Technology-Facilitated Gender-Based Violence (TFGBV) on Survivors in Jordan*, Final-JOSA-KVINFO-TFGBV-Study-1.pdf

[21] Renaissance Strategic Centre (2026) *How TFGBV Impacts the Democratic Participation of Young Women: Jordan as a Case Study*, TFGBV-report.pdf

[22] International Media Support (IMS), *Online gendered disinformation and sexist hate speech*, IMS-Online-gendered-disinformation_final.pdf

[23] Renaissance Strategic Centre (2026) *How TFGBV Impacts the Democratic Participation of Young Women: Jordan as a Case Study,* TFGBV-report.pdf, p.2

[24] Analyse & Tal, KVINFO and TrygFonden (2025) *Digital hverdagssexisme: Hvordan forebygger og håndterer man den hårde tone i onlinedebatter*, https://kvinfo.dk/wp-content/uploads/2025/04/Digital-hverdagssexisme_20250430.pdf, the quote is translated from Danish using ChatGPT

identity, and technology intersect, and how digital misogyny and harm translates into physical-world consequences. For CSOs, journalists and researchers, who are targeted by this AI-model, this understanding is a necessary foundation for data processing and analysis.

3. THE AI TOOL FOR DETECTION AND RESEARCH

The increasing scale, speed, and complexity of online hate speech means that manual observation and reporting, while valuable, cannot keep pace with the content generated across digital platforms. In response to this challenge, AI–powered tools have emerged as important resources for detecting, analysing and researching gendered online hate speech and TFGBV. These tools offer new possibilities to systematically document digital harm, generate evidence, improve protection mechanisms and support advocacy and policy engagement.

CONCEPTUAL FRAMEWORK

Tadwein developed the conceptual framework of this AI model, in collaboration with KVINFO. Drawing on feminist research, international normative frameworks and localized evidence, ensuring that the conceptual foundation of the project was both theoretically sound, localised and contextually relevant.

The term Technology-Facilitated Gender-Based Violence was selected, over more commonly used but narrower terms such as “online violence,” “cyberbullying,” or “digital abuse”. TFGBV is a gendered and intersectional form of violence, emphasizing how digital harm disproportionately affect women, girls and sexual minorities, and how this harm is shaped by intersecting factors such as age, class, sexuality, gender identity, ability, and social status. This conceptual clarity was critical to ensuring that the AI model does not treat online abuse as isolated or neutral behavior, but rather as part of broader systems of gender-based oppression.

DEVELOPING THE TOOL

During the project, based on experiences from working with Jordanian Arabic, JOSA adopted a machine learning lifecycle that consists of four phases – discovery, training, deployment, and monitoring. In the discovery phase, the problem is defined, data is gathered, and goals are set. The training phase involves building and training the model, using the collected and classified data. Deployment integrates the model into the production environment. Monitoring continuously assesses the model's performance and makes necessary adjustments.[25]

This AI-model has been developed based on large amounts (tens of thousands of comments) of humanly annotated data, by Egyptian Arabic speaking annotators. The model is also

[25] Democracy Reporting International (2024) *Online Public Discourse in the MENA Region, Persistent Tactics of Disinformation and an Increase in Online Gender-Based Violence*, 65b21f4539b6e.pdf p.65-66 (NB most of above is a quote)

programmed to detect certain keywords and linguistic structures, which means that a hybrid approach has been used to develop the tool. The training continued until a satisfactory accuracy level was reached.

Tadwein, as an organisation strongly rooted in Egyptian feminist discourse, provided conceptual guidance on how to distinguish harmful, gendered violence from legitimate speech, helping to mitigate risks of overreach, misclassification, or political misuse of the AI tool. This distinction is essential both ethically and technically, particularly given the restrictive civic space and the history of state surveillance and prosecution of marginalised groups. Tadwein further strengthened the knowledge base by embedding Egyptian socio-political and legal context into the conceptual framework.

Practically, Tadwein monitored 75 individuals, initiatives, and pages across social media platforms, including Facebook, Instagram and TikTok, to identify potential hate speech and TFGBV-related content. This monitoring was conducted in alignment with the criteria outlined in the project methodology. For each selected account, Tadwein provided analytical comments, direct URLs to the accounts, and links to relevant posts identified for further review and analysis.

In parallel, Tadwein reviewed the methodology section in detail and systematically flagged unclear or ambiguous points that could potentially mislead the monitoring and data annotation processes. These observations were shared to support methodological clarity, consistency, and accuracy throughout the implementation of the monitoring and annotation phases. Furthermore, Tadwein worked closely with JOSA to review the accounts monitored by JOSA, ensuring that the selected accounts and content were consistent with the Egyptian context and fully aligned with the agreed methodology.

RELEVANT CONTEXTS

The AI-based tools are valuable in a research, media and civil society contexts. They enable organisations to monitor large volumes of public online content in real time or retrospectively, identify spikes in abusive language, and uncover recurring narratives or coordinated harassment campaigns. AI-tools can also support comparative analysis across platforms, time periods, or geographic areas, making it possible to document trends and patterns that would otherwise remain invisible. For advocacy purposes, the data generated through these tools can strengthen evidence-based engagement with policymakers, social media platforms, and international human rights mechanisms.

A MORE RELIABLE AND ACCESSIBLE TOOL

While many commercially available AI tools offer limited insight into how decisions are made or how models are trained, this model gives users full control over which data is analysed, while knowing that the training and development process has followed ethical standards for data management, which increases trust in research findings. This is especially important in engagement with authorities, and for public accountability towards the CSO sector, media and academia.

Another aspect of great importance in this project has been the usability for non-technical users. The interface is intuitive and easy to use, while it is open source, which gives additional

opportunities for tech-savvy users.[26] This means that the model is quite versatile, but with emphasis on useability for CSOs with less technical capacity.

When combined with contextual and other crucial knowledge, this AI-powered detection tool will be able to significantly enhance efforts to document online gendered hate speech and TFGBV.

You can access the Nuha platform and explore how to use it here: Nuha

4. USABILITY AND TECHNICAL ACCESSIBILITY

A central component of the project was to ensure the AI-powered research tool would be useful and accessible for civil society organisations of different capacity, and particularly those working in the field of gender justice and women's rights. Many CSOs across the MENA region operate with limited technical resources and budgets, and small teams. Egypt is unfortunately no exception. For this reason, it was essential for the tool to be intuitive and easy to integrate into the current work of organisations, without requiring specialised technical expertise. Meanwhile it needed all the features expected of university and other researchers. The usability therefore focused not only on the tool's technical performance but also on how easily it could be integrated into the work of practitioners and researchers in and on Egypt.

The interface was designed to be clear and straightforward, allowing users to upload large datasets, run automated analyses, and view results through different visualisations. After the introductory workshops, most participating organisations were able to navigate the system with minimal instruction, demonstrating that AI-assisted monitoring can be made accessible even to teams with limited backgrounds in data science or machine learning. Having trained users in the country functions as an additional support, also beyond the time frames for the project.

TRAINING FOR APPLYING THE TOOL

The introductory workshops played a crucial role in enabling organisations to make full use of the tool. They introduced the basic concepts of AI-based detection, explained how the model processes text (and even emojis), and outlined the limitations and appropriate uses of automated analysis. Participants were encouraged to approach the tool not as a standalone solution but as an extension of and complement to their existing expertise. These sessions helped demystify the AI-tool, and gave users confidence in applying it to their own activities. The workshop format also enabled CSO representatives and researchers to share their experiences, highlight contextual challenges, and collectively troubleshoot common issues such as ambiguous language, irony etc.

Two two-day training sessions were conducted in Cairo in December 2025, targeting CSOs, researchers, youth-led initiatives and journalists. The trainings aimed to introduce the Nuha AI model, explain its purpose and functionality, and demonstrate how Egyptian CSOs can use the

[26] Project reference group input, September 2025

tool to detect hate speech and TFGBV on social media platforms. A total of 40 participants took part in the trainings. The sessions were highly practical and interactive, allowing participants to test the AI model directly and engage in discussions on its potential applications, limitations, and areas for further improvement.

The tool can be operated through a standard web browser, without requiring installation of complex software or access to high-performance computing resources. This simplicity means that organisations with basic office equipment and moderate internet connectivity will be able to fully engage with the tool. Importantly, the tool allows users to work at their own pace and adapt the workflow to suit their operational realities, whether they are conducting continuous monitoring or analysing specific events, profiles or campaigns.

THE WEBSITE: FOR RESEARCH AND PUBLICATIONS

The website is also an opportunity to showcase organisations in Egypt that address TFGBV in different ways, including their publications and research on the topic, in both Arabic and English. The aim is for the website to strengthen links between the Nuha tool and existing feminist research and advocacy efforts in Egypt.

Adding large scale data to case studies, or conducting smaller studies (focusing on certain hashtags/ profiles/ time periods/ areas) is now accessible also for smaller organisations. This means that we hope to see diverse research reports building the case towards prevention and protection from gendered online hate speech and TFGBV. Also, for those already interested in the topic, instead of manually sifting through thousands of social media posts, can focus their efforts on interpreting results, identifying trends, and preparing advocacy materials. It could also help reveal patterns that have previously been suspected but not proven with data. For example, analysing trends around elections or other events or incidents.

The trained users will continue to support newer users, and to introduce the tool, both through campaigns and more hands-on support.

5. CHALLENGES AND LIMITATIONS

The use and misuse of AI currently takes up a lot of band width in public debate. There are significant challenges as well as risks associated with the use if AI, not least in any kind of human rights setting. The fact that this tool will only process publicly available information is important. When using the model, the data harvesting is semi-manual, which limits the risks relating to automated tracking of online behaviour, which can be related to surveillance, profiling, or the targeting of vulnerable groups—particularly in regions where digital rights protection and regulation are weak. Collecting only publicly available data ensures compliance with human rights principles but also limits the scope of analysis when abusive content occurs in semi-private or closed online spaces.

SAFETY

This also means that using the AI-tool will not entail processing of sensitive or personal information. If such information should already be made public, it could end up in the tool, in

which case it is important to determine practices for legal action, and the responsibility of the researcher or activist that harvests the data if such sensitive or leaked information is discovered.

Handling large amounts of material, including material with threats, harassment and violence is inevitable when using the model. This requires careful data protection to avoid exposing victims/survivors of online abuse to additional risks. There is a need for secure infrastructure, clear data management policies, as well as robust anonymisation.

Data protection is also related to the need, or encouragement, of securing personal support systems for researchers and activists who will work with the Egyptian Arabic AI-model. Some, if not a vast deal, of the material that they will process while working on this topic will be violent, and research shows that hateful information affects us, even though the messages are not directed towards us.[27] Therefore, it is important both to limit the number of persons viewing the material, and to support the ones who do.

BIAS, IRONY AND DIALECTS

One of the most persistent technical limitations encountered during the data annotation phase was the difficulty in accurately interpreting the complexity of human language. Online communication is rarely straightforward; it involves sarcasm, ironic and coded language, humour, and contextual references that automated systems need additional efforts to be trained on, in order to categorise correctly. The lessons learnt from the pilot project and model with Jordanian Arab dialect was crucial is this case. Despite these challenges, the Egyptian Arabic AI-tool reached an impressive accuracy during the different phases of training. It remains important however, to maintain human oversight of any analysis. Also, it is crucial to present the methodology used in a transparent manner, so that the reader fully understands the accuracy of the data.

A related challenge concerns the bias within AI models. Machine learning models are only as reliable as the data on which they are trained, and as we all know, we live in a society of bias. This is also why gender-based hate speech has been known to be more difficult to detect than for example racist hate speech.[28] Gender biases are still internalised in our societies, meaning that they are oftentimes unconscious. During this project, biases were countered by the extensive use of human annotators. This is also why having a feminist awareness was crucial throughout the development process – both among the technical staff and annotators.

CONCERNS

The negative environmental and climate aspect of AI is possibly less talked about than ethical considerations. Fact remains that training an AI consumes large amounts of water (for cooling in data centres) as well as electricity, and electronic hardware. In the case of this AI-model, the day-to-day use will be less resource demanding than large-scale AI language models, due to the much lower number of daily users.

---

[27] See for example: Vicarious trauma after viewing media - Wikipedia

[28] Analyse & Tal, KVINFO and TrygFonden (2025) *Digital hverdagssexisme: Hvordan forebygger og håndterer man den hårde tone i onlinedebatter*, https://kvinfo.dk/wp-content/uploads/2025/04/Digital-hverdagssexisme_20250430.pdf

Finally, it is important to not place too much faith in AI. This model will be able to help detect hate speech and TFGBV at a larger scale than manual screening, but it cannot address the underlying cultural, political, and structural drivers that fuel misogyny. It is one aspect of a the complex, interconnected approaches needed to improve gender justice. It is important to use the data to advance our knowledge and understanding of gendered hate speech online and of TFGBV, and that this knowledge is used in advocacy, policy making and practices, to create the changes we aim for.

6. RECOMMENDATIONS

Automated systems can rapidly process large volumes of content and identify overt patterns, but they lack the cultural, social, and contextual understanding necessary to fully interpret nuanced or coded forms of abuse. Therefore, core recommendations include to:

- approach AI as a complementary tool to human expertise. Ideally, researchers, CSOs, activists and others using the developed model should adopt a hybrid workflow that combines the AI-detection with human review. By integrating current knowledge and expertise, lived experiences of specifically exposed online users and contextual understanding, findings will be both accurate and meaningful for advocacy, prevention and protection.
- build and develop one's own capacity, both on how to operate the tool, and on an understanding of the model, together with critical analysis. The intention is that the model should be useful for strategic monitoring, including identifying patterns, and online trends, for some time to come. Using this AI-tool has already been facilitated by several trainings for researchers, journalists, CSO representatives and activists.
- Promote the tool to public institutions in Egypt, who could also benefit from using it.
- consider how to keep the data model updated, to ensure accurate classification and processing. Possibly there is interest at a university or similar to "inherit" the model, provided it will continue to be open source, for the purpose of keeping it fully up to date, seeing as online language, slang, memes and expressions change rapidly.
- for anyone working with large scale data harvesting to be required to develop internal protocols for data management, including secure handling of sensitive content.

AI-generated evidence can also be used as a foundation for advocacy aimed at improving platform accountability. It could potentially be used to pressure social media companies to strengthen moderation policies, respond more quickly to harmful content, and other actions that are known to be conducive to more constructive online interactions. As mentioned earlier, the goal is that the data and knowledge built should be used towards changes and improvements in policy and practice, via for example evidence-based advocacy.

Finally, policy makers should hear journalists and CSOs, since they are often at the frontline when it comes to detecting new forms of TFGBV or gendered online hate and harassment. As digital communication evolves, so too will the linguistic forms and strategies of online hate. Future-proofing monitoring efforts requires continuous learning, updating tools, and experimenting with new methods. Strengthening regional cooperation is crucial, as shared

datasets, cross-country learning, and collaborative analysis can significantly enhance joint efforts. By building networks and fostering knowledge exchange, CSOs and journalists can collectively increase their impact and develop a more unified regional response to online gender-based violence.

7. CONCLUSION: A SUPPORTIVE TOOL FOR GENDER JUSTICE

The experiences gained through the AI-powered monitoring project demonstrate both the potential and the complexity of using technology to address online gendered hate speech. It gives new possibilities to discern patterns and trends, currently and over time, with new possibilities for more and stronger voices in evidence-based advocacy towards legal and practical changes, as well as for awareness campaigns and protection, in a broad perspective. Evidence generated through AI-assisted research can strengthen engagement with social media platforms, policymakers, and international human rights mechanisms, helping to promote transparency, accountability, and the protection of digital rights for all of us.

The work also underscores the importance of adopting (and challenges with) a contextually grounded approach to new methods of detecting hate speech and TFGBV. AI is most effective when treated as a supportive tool rather than a standalone solution.

Looking ahead, the project offers important insights and experiences for the development of similar tools in other countries or dialects. Using similar methodology in several countries would offer unique comparability, for additional, regionally relevant research and advocacy.

By equipping organisations, journalists and researchers with AI-powered tool, as part of a comprehensive range of methodologies and practices, this project will contribute to broader public and political participation and interaction, in online spaces – which have the potential of being both democratic and accessible. This can serve as a model for other contexts, in our joint, global efforts, to eradicate online gender-based violence, to the benefit of increased political participation, democracy and gender justice.

8. SOURCES

Analyse & Tal, KVINFO and TrygFonden (2021) *Angreb & had i den offentlige debat på Facebook*. Available at: Angreb-og-had-i-den-offentlige-debat-paa-Facebook.pdf

Analyse & Tal, KVINFO and TrygFonden (2025) *Digital hverdagssexisme: Hvordan forebygger og håndterer man den hårde tone i onlinedebatter*. Available at: https://kvinfo.dk/wp-content/uploads/2025/04/Digital-hverdagssexisme_20250430.pdf

Danish Institute for Human Rights (DIHR) (2024) *DEBAT: Demokratiet lider, når kvinder flygter fra den offentlige debat*. Available at: DEBAT: Demokratiet lider, når kvinder flygter fra den offentlige debat | Institut for Menneskerettigheder

Democracy Reporting International (2024) *Online Public Discourse in the MENA Region, Persistent Tactics of Disinformation and an Increase in Online Gender-Based Violence*. Available at: 65b21f4539b6e.pdf

EuroMed Rights *Online gender-based violence: What scenario for the MENA region?* Available at: Home - EuroMed Rights

International Center for Journalists (ICFJ) *Big Data Case Study: Online Violence Against Journalist Ghada Oueiss*. Available at: ICFJ_BigData_Ghada Oueiss_Online Violence.pdf

International Media Support (IMS) *Online gendered disinformation and sexist hate speech*. Available at: IMS-Online-gendered-disinformation_final.pdf

Jordan Open Source Association (JOSA) and KVINFO *The Impact of Technology-Facilitated Gender-Based Violence (TFGBV) on Survivors in Jordan*. Available at: Final-JOSA-KVINFO-TFGBV-Study-1.pdf

Posetti, J. et al. (2025) *Tipping point: The chilling escalation of online violence against women in the public sphere in the age of AI*. UN Women. Available at: tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

Renaissance Strategic Centre (2026) *How TFGBV Impacts the Democratic Participation of Young Women: Jordan as a Case Study*. Available at: TFGBV-report.pdf

Special Rapporteur on the promotion and protection of freedom of opinion and expression (2021) *Gender justice and freedom of expression*. UN General Assembly, A/76/258. Available at: A/76/258: Gender justice and freedom of expression - Report of Special Rapporteur on the promotion and protection of freedom of opinion and expression | OHCHR

Storey, J. E. et al. (2021) 'Technology-Facilitated Intimate Partner Violence: An Examination of Prevalence, Perpetration Type and Methods and the Impact of COVID-19', *Journal of Interpersonal Violence,* Technology-Facilitated Intimate Partner Violence: An Examination of Prevalence, Perpetration Type and Methods and the Impact of COVID-19 - PubMed

UN Women (2021) *Violence against women in the online space: insights from a multi-country study in the Arab States*. Available at: Summary_Keyfindings_Final_EN.pdf

UN Women (2023) *Expert Group Meeting Report: Technology-facilitated violence against women*. Available at: Expert-Group-Meeting-report-Technology-facilitated-violence-against-women-en.pdf

UN Women and UNDP *Technology Facilitated Gender-Based Violence: Developing a shared research agenda*. Available at: technology-facilitated-gender-based-violence-shared-research-agenda-en.pdf

UNFPA *Technology-facilitated Gender-based Violence (TFGBV)*. Available at: Technology-facilitated Gender-based Violence: A Growing Threat | United Nations Population Fund

United Nations Secretary-General (2024) *Intensification of efforts to eliminate all forms of violence against women and girls: technology-facilitated violence against women and girls*. A/79/500. Available at: a-79-500-sg-report-ending-violence-against-women-and-girls-2024-en.pdf

Zuleta, L. et al. (2024) *Ytringsfrihed og selvcensur*. Danish Institute for Human Rights (DIHR). Available at: Ytringsfrihed-og-selvcensur_DK_juni2024.pdf

THE TOOL

You can access the Nuha platform and explore how to use it at this website: Nuha

ABOUT THE ORGANISATIONS

International Media Support (IMS) works on one of today's defining challenges: ensuring that people everywhere can access trustworthy, fact-based information. The way societies understand the world - and shape their future - depends on who produces information, how it circulates, and whether it can be trusted. IMS exists to protect and strengthen the information ecosystems people rely on. We work with bold and brave local partners in contexts affected by conflict and social tension, helping ensure that ethical, public-interest information can be produced and shared safely. Through long-term, locally grounded partnerships, we co-create practical and innovative solutions - and respond rapidly in moments of crisis or opportunity.

KVINFO is Denmark's knowledge center for gender and equality. We collect and produce knowledge, develop tools and work for solutions to specific challenges about gender and equality in Denmark and internationally. KVINFO works at the crossroads between knowledge, politics and practice. We create an overview of research and knowledge on the field and bring this knowledge into play broadly to decision-makers, media, companies, organisations and the public in Denmark and internationally. We apply knowledge to drive change through tools and solutions tailored to gender-specific challenges. We are committed to creating public value, for example, pointing out and responding to significant societal challenges directly related to gender and equality or where knowledge about gender is a key to creating change.

The Jordan Open Source Association (JOSA) is a non-profit organisation based in Amman, Jordan. The association is among the few non-profits registered under the Jordan Ministry of Digital Economy and Entrepreneurship. JOSA's mission is to promote openness in technology and to defend the rights of technology users in Jordan. We believe that information that is non-personal – whether it's software code, hardware design blueprints, data, network protocols and architecture, content – should be free for everyone to view, use, share, and modify. Our belief also holds that information that is personal should be protected within legal and technological frameworks. Access to the modern Web should likewise remain open.

Tadwein for Gender Studies was established in 2014 with the aim of promoting evidence-based awareness on gender issues, implementing projects, formulating policies, and taking necessary actions to enhance the status of women in Egyptian society and to reduce violence against women and girls in general.

ANNEX 1 – KVINFO'S STRATEGY ON TFGBV

ADVANCING EQUALITY IN PUBLIC AND POLITICAL INTERACTIONS BY PREVENTING AND RESPONDING TO TECHNOLOGY-FACILITATED GENDER-BASED VIOLENCE

VISION

Everyone, including human rights defenders, activists, journalists, politicians, researchers irrespective of gender, age, sexuality, ethnicity and other intersecting factors are able to express their views, influence politics and public debate in online and offline spaces, without risk of TFGBV.

MISSION

In the countries where KVINFO works, we seek to reduce the risk and consequences of TFGBV at structural, individual and norm-critical levels:

- KVINFO *builds knowledge* on TFGBV and its consequences for equality in public and political interactions, and for improved practices in prevention and response
- KVINFO *supports civil society organisations and through them relevant duty bearers* to improve
    - prevention and response to TFGBV enabling the protection and support to victims/survivors
    - legal frameworks and institutional responses through advocacy and awareness efforts, fostering legal protection and resilience among users in cyberspace, especially youth

AIM OF THIS STRATEGY

TFGBV is prevented and responded to among youth, women and other groups in higher risk, such as women human rights defenders, women politicians, feminist activists and journalists, thereby enhancing gender just political representation and interaction in civil society, in representative fora and public debate.

DEFINITION OF TFGBV

KVINFO applies this definition of TFGBV:

"Technology-Facilitated Gender-Based Violence, or TFGBV, is an act of violence perpetrated by one or more individuals that is committed, assisted, aggravated and amplified in part or fully by the use of information and communication technology (ICT) or digital media, against a person on the basis of their gender."[29]

KVINFO applies this definition in order to emphasize how digital gender-based violence results, or is likely to result, in physical, sexual, psychological, social, political, and/or economic harm, or other infringements of rights and freedoms, leading to withdrawal from public and political participation, thus causing a democratic deficit.

[29] Coined by UNFPA, see (https://www.unfpa.org/TFGBV), see also UN Women for TF-VAW, "Technology-facilitated violence against women" (https://www.unwomen.org/en/digital-library/publications/2023/03/expert-group-meeting-report-technology-facilitated-violence-against-women)

BACKGROUND ANALYSIS

Online spaces have become an integral part of public life, and offer possibilities for immediate expressions of individual and collective opinions, demands, aims and aspirations, presenting opportunities for participation in public and political debates and activism. This development, however, has highly ambiguous consequences for gender justice, including the right to equal public and political participation and a life free from threats and violence.

Online participation and expression frequently lead to attacks that usurp gender stereotypes, based on or combined with derogative references to gender identities. These acts of gender-based violence in cyberspace, or TFGBV, signifies the harassment, hate speech, defamation, slandering and unwanted exposure in cyberspace etc. of individuals and groups.

TFGBV affects millions of people, and the impact of TFGBV on women, girls and others, including LGBTIQ+ persons is profound and multifaceted. Sixty percent of women in the Arab states reported experiencing some form of online harassment in the preceding year.[30] In a similar manner, more than half of all women in Caucasus and Eastern Europe experienced at least one form of TFGBV,[31] whereas globally, different studies indicate that between 16-60% of women are affected.[32] Perpetrators are often associated with anti-gender and anti-feminist individuals or groups, but can also be state agents.

Victims/survivors often experience severe consequences, including anxiety, depression, and post-traumatic stress disorder that frequently causes withdrawal from political and public participation in online as well as offline spaces.

Increasing violence against women in politics

As documented by research, violence against women in politics is increasing at global level. TFGBV sustains this trend, thus exacerbating an already existing democratic deficit. Often committed in a continuum between online and offline spaces, the consequences of violations that are initiated in digital environments are difficult to distinguish from offline realities, and vice versa. Moreover, the relationship between TFGBV and femicide is increasingly troubling. Over 50% of Arab women journalists said they had been attacked or abused offline in incidents seeded online.[33]

This means that TFGBV poses a grave threat to not only well-being, safety and freedom of expression, but also to public participation and political influence. Evidence suggests that certain actors, especially those exposed in public are more at risk, such as activists – feminists, women human rights-, land rights and environmental defenders, also academics, advocates, and politicians and young people in general.

---

[30] UN Women, 2021: Violence against women in the online space: insights from a multi-country study in the Arab States. https://arabstates.unwomen.org/en/digital-library/publications/2021/11/violence-against-women-in-the-online-space#view

[31] UN Women, 2023: The dark side of digitalization: Technology-facilitated violence against women in Eastern Europe and Central Asia, https://eca.unwomen.org/en/digital-library/publications/2023/11/the-dark-side-of-digitalization-technology-facilitated-violence-against-women-in-eastern-europe-and-central-asia

[32] United Nations, Report of the General Secretary, 2024: Intensification of efforts to eliminate all forms of violence against women and girls: technology-facilitated violence against women and girls, *available at*: a-79-500-sg-report-ending-violence-against-women-and-girls-2024-en.pdf

[33] UNESCO, 2022 The Chilling: global trends in online violence against women journalists; research discussion paper - UNESCO Digital Library

ANNEX 2 – IMS' LEARNING BRIEFS

Online gendered disinformation and sexist hate speech IMS-Online-gendered-disinformation_final.pdf

Current digital infrastructures present threats to gender equality, democracy, peace and the positive impacts – in media and societies around the world – accomplished in the past 30 years. Issues like online gendered disinformation and sexist hate speech are growing at an alarming rate, and the consequences
are colossal.

This learning brief focuses on the issues of online gendered disinformation and sexist hate speech against women, girls and non-binary people who work or appear in the media and what media development organisations can do to address them.

Safety of women journalists IMS-Learning-Brief-1-Safety-of-women-journalists-1.pdf

In the past decade, awareness that women in the media are subject to gender-based violence has grown as a number of ground-breaking reports have been published, establishing that violence and threats against women journalists have reached endemic proportions. Three out of four women journalists have now been subject to online violence, and the killings of women journalists have increased at an unprecedented speed. Along with this awareness has come a greater understanding that these threats pose a serious challenge to media freedom and development.

Addressing the safety of women journalists is not only a matter of protecting individuals but also a means of safeguarding democratic values, human rights and the richness of media representation. It is essential for fostering an environment where all journalists can work freely, contribute to informed public discourse and play a vital role in shaping the societies they serve.

KVINFO
CHRISTIANS BRYGGE 3
1219 KØBENHAVN K
TEL +45 33 13 50 88
kvinfo@kvinfo.dk
www.kvinfo.dk