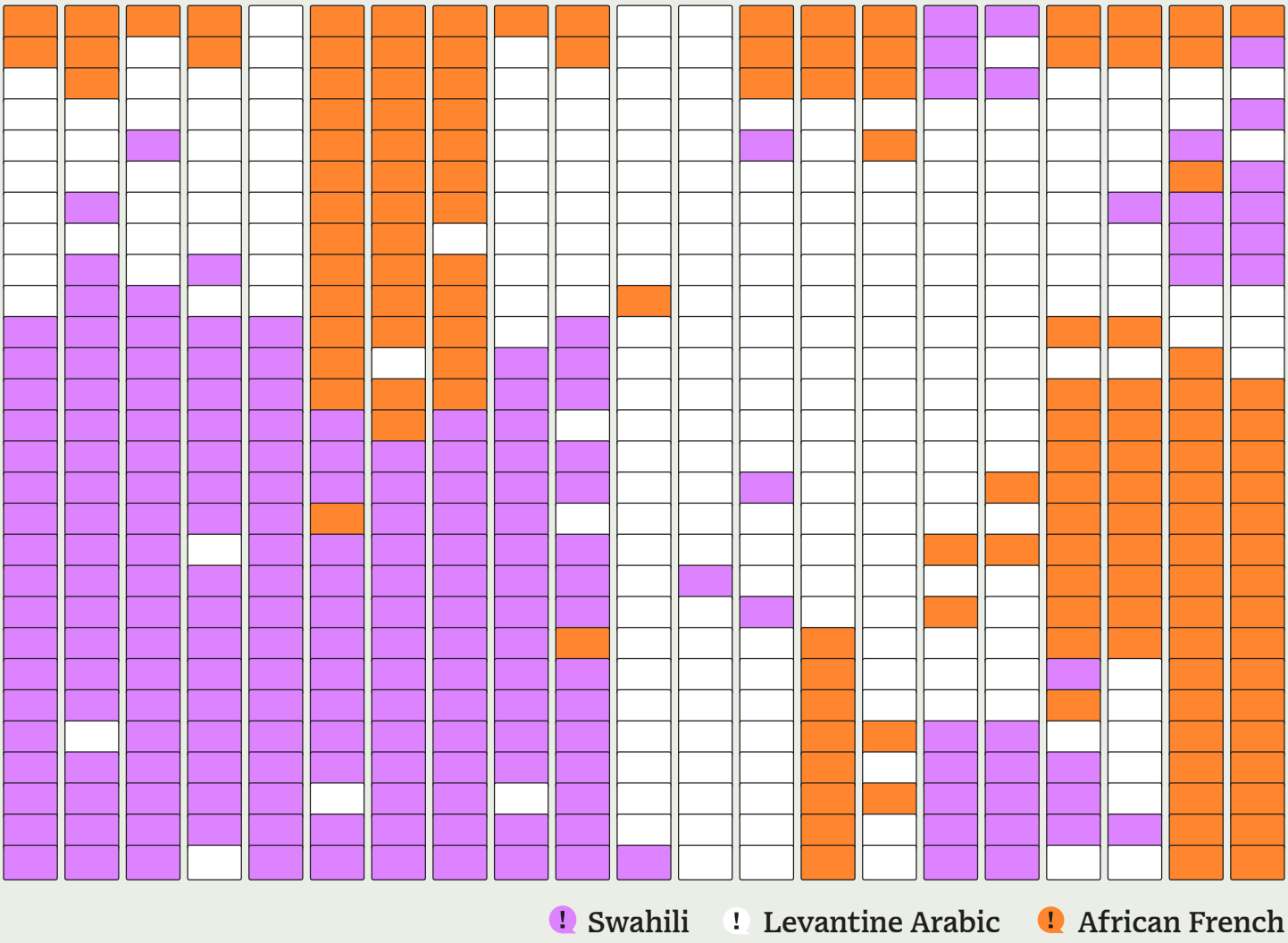


Detecting tech-facilitated gender-based violence in African French, Levantine Arabic, and Swahili

September, 2026

meedan.org

Meedan

Executive Summary

From sexual harassment to doxxing to other coordinated forms of abuse, technology-facilitated gender-based violence (TFGBV) has been well documented across major social media platforms over the past decade. Yet the technological interventions designed to tackle this problem have focused overwhelmingly on English and other European-language content in North America and Europe, leaving most of the world with far weaker online protections.

Over the past three years, this inequity has driven Meedan's program and research teams to study the unique linguistic and contextual dynamics of TFGBV in the Larger World.

Building from our 2024 exploration of gendered disinformation in South Asia, in 2025 we developed a new study tracking and analyzing TFGBV in African French, Levantine Arabic, and Swahili on the social web. This mixed-methods research project was intended to deepen public understanding of how TFGBV manifests in major languages of Western Asia and Africa. The research also contributed to improving machine learning solutions that can help reduce the spread of TFGBV.

Led by Meedan, a technology nonprofit, the project's regional collaborators included the Kenya ICT Action Network (kictanet.or.ke), the Jordan Open Source Association (josa.ngo), and Her Safe Space (thebrainbuilders.org/hersafespace/) which serves communities across West Africa.

This work forms part of Catalyst, a scalable 18-month consortium project tackling online misogyny and gender-based violence as pathways to radicalisation and violent extremism. Catalyst is led by the Christchurch Call. We received further support from The United Nations Population Fund (UNFPA) and The Swedish International Development Cooperation Agency (SIDA).

A preprint of our academic research paper from the project is available online.[1]

1 A preprint of our academic research paper is available here: https://drive.google.com/file/d/1dP8hjIH_d8GnyMtr-mAylH9E1-56qqCh/

What is technology-facilitated gender-based violence?

Technology-facilitated gender-based violence refers to acts of gender-based harm carried out, amplified, or enabled through digital technologies and online platforms. It includes behaviors such as harassment, hate speech, threats, disinformation, non-consensual image sharing (including AI-generated content), doxxing, and coordinated abuse, disproportionately affecting women, gender nonconforming individuals, and other marginalized groups.

TFGBV[2] is a global problem consistently linked with negative outcomes, including depression, anxiety, and stress (Benítez-Hidalgo et al., 2026) and often affects minoritized populations more acutely than others (Turner et al., 2024). In a recent survey of women in the public sphere from 119 countries, 41% of respondents reported that they had experienced offline attacks and harassment linked to online attacks (Posetti et al., 2025). Beyond its immediate online impacts, TFGBV can silence participation, restrict freedom of expression, and reinforce exclusion from digital public spaces, while also contributing to offline psychological and physical harm, other safety risks, and broader social and political marginalization.

Unique contours of TFGBV in the Larger World

In the Larger World, TFGBV is often missed by detection systems that do not account for low-resource languages or contexts, or the lived experiences of those who most commonly experience this type of harm.[3] These gaps make it harder to effectively detect and respond to such content, and to prevent harm.

Online platforms often deploy a special type of algorithm, known as a classifier, to detect and help reduce the prevalence of harmful speech. Classifiers are trained to identify certain types of content, such as hate speech, typically by processing hundreds of examples of that type of content. Once sufficiently trained, a classifier can interpret the meaning of a message or image on social media and will then assign a label based on that meaning (for example, it will label a piece of data as either “harmful” or “not harmful.”) While many classifiers can identify and help reduce TFGBV, prior research has shown that these classifiers tend to perform well on English-language test sets, but that their performance drops substantially on non-English and real-world data (Tonneau et al., 2025). This gap not only causes immediate harm to at-risk groups but can also contribute to broader offline safety and security risks (Riedl et al., 2024).

> **Hypothesis: If we train a classifier using data gathered by communities that are encountering TFGBV first-hand, the classifier will improve.**

2 We use TFGBV as an umbrella term that includes misogyny, technology-facilitated sexual violence, online sexual harassment, and other similar concepts.

3 Low-resourcedness describes languages which are underrepresented in natural language processing (NLP) resources by a lack of datasets and dedication of research resources.

Our objectives

The overarching objective of this project was to improve existing off-the-shelf classifiers by introducing a new training dataset for TFGBV detection in African French, Levantine Arabic, and Swahili. We posited that if the datasets used to train classifiers better reflected the experiences of communities most affected by online harm, they would improve classifiers' ability to detect TFGBV in these languages and regions. In turn, this would enable platforms and online communities to reduce TFGBV prevalence in these languages and allow researchers to study its proliferation more effectively. With this in mind, we sought to:

- Develop and deploy a methodology for community-led data collection in collaboration with our three regional partner organizations in Francophone West Africa, Jordan, and Kenya;
- Establish TFGBV detection benchmarks on existing classifiers for African French, Levantine Arabic, and Swahili;
- Develop classifiers for these three languages that surpass existing benchmarks; and
- Release the datasets and classifiers for other researchers and technology platforms to use and further refine.

A community-first approach

Our regional partnerships were a cornerstone of this project's design and implementation. We prioritized working with organizations with prior expertise in TFGBV and a core focus on serving communities that speak African French, Levantine Arabic, and Swahili. Selection criteria included a commitment to utilizing technological tools for data gathering and, reflecting the project's core focus, leadership by women or non-binary individuals. The following organizations were selected as our partners:

- Jordan Open Source Association (josa.ngo), a non-profit organization based in Amman that works at the intersection of open technology, digital rights, and public policy in Jordan and the wider region, led data collection for Levantine Arabic.
- Kenya ICT Action Network (kictanet.or.ke), a multistakeholder think tank that brings together government, civil society, the technical community, academia, and industry to shape ICT policy and regulation in Kenya, led data collection for Swahili.
- Her Safe Space (thebrainbuilders.org/hersafespace) is part of Brain Builders Youth Development Initiative (BBYDI), a non-profit, youth-led civic organization in Nigeria that focuses on promoting good governance, civic engagement, and sustainable development by empowering young people and communities through education, advocacy, and leadership programs. Her Safe Space led data collection for African French. While BBYDI is based in Nigeria, the organization works across Francophone Africa, in countries including Togo, Cameroon, and Benin.

Data collection & annotation

Our partners' primary role in the project was to collect social media posts containing context-specific instances of TFGBV. Teams then annotated those posts using a tiered taxonomy covering specific types of TFGBV, such as sexual harassment, threats and incitement of physical violence, intimate image abuse, and various others. We trained two native speakers from each partner organization in the annotation technique and paired this training with a trauma mitigation workshop designed to help annotators anticipate and cope with secondary trauma that can result from exposure to harmful online content.

We leveraged our combined technical and programmatic expertise to crowdsource local, community-reported instances of TFGBV from each of our regional partners' online communities. Each partner established and promoted tiplines on WhatsApp and Facebook using Meedan's Check software. Partners used tiplines to solicit examples and reports from their communities, allowing the team to capture the lived experience of people in the regions. We also adopted keyword-based sampling on social media to boost the number of relevant examples. The resulting full dataset is called "**ALTO**: **A**frican and **L**evantine **T**ech-facilitated Gender-based Violence C**o**rpus."[4]

Benchmarking existing tools

Meedan's research team began this phase by seeking to identify all existing, publicly available TFGBV datasets in African French, Levantine Arabic, and Swahili. We found some resources that focused on abusive and hate speech detection writ large, but relatively few of these focused specifically on gendered harm (Vidgen & Derczynski, 2020). This phase of the project underscored the noticeable industry-wide gap in resources needed to effectively support TFGBV detection and machine learning applications.

We used these datasets to benchmark the performance of gendered harm detection by existing open-source trust and safety and content moderation models across the languages we studied. We also identified four off-the-shelf models from leading developers to benchmark against the existing datasets. These models included CoPE-B-A4B (from Zentropi AI), Perspective API (from Google), Learning From the Worst (from Meta), and GPT-OSS-Safeguard 20- and 120-billion-parameter models (from OpenAI).

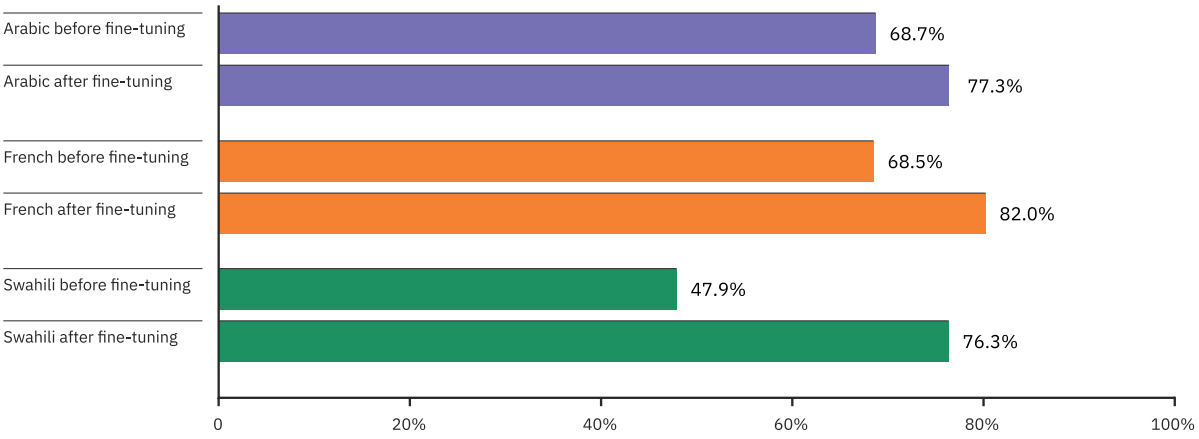
With these resources in hand, we were able to test the extent to which we could improve existing classifiers using our newly crowd-sourced datasets.

> **Conclusion: When a community is meaningfully engaged in building a technology, that technology will be more likely to serve its members effectively.**

Results

When we incorporated the crowd-sourced ALTO dataset, we found that our hypothesis was correct. Once trained on our dataset, models showed improved performance at detecting TFGBV in the target languages, and their performance exceeded that of most off-the-shelf models. Gendered harm remains a challenging category for most current models to detect across the three languages, with consistently lower recall. Dominant, off-the-shelf models on average detected 68.7% of TFGBV in Arabic, 68.5% in French, and 47.9% in Swahili before we fine-tuned them on our dataset. After fine-tuning on ALTO, the performance increased to 76.4% for Arabic, 80.2% for French, and 76.3% for Swahili. Our datasets yielded greater improvements for Arabic and French (despite our having smaller datasets for those languages) when trained monolingually than for Swahili, which required multilingual training (i.e. training on all three languages together) to see improvements. We speculate that this is due to the fact that there are fewer training datasets in the Swahili language model than there are for the other two languages. This indicates a need for greater investment in Swahili datasets.

4 Dataset and trained models can be found here: https://huggingface.co/collections/meedan/alto



Model performance in TFGBV detection before and after fine-tuning on ALTO

Overall, these findings show that current off-the-shelf models cannot reliably capture TFGBV content in the languages studied. Further language and harm-specific resources are required from researchers and model developers. The full results of this study can be found in our academic paper.[5]

Conclusion

At a high level, our study demonstrates that when a community is meaningfully engaged in building a technology, the technology is more likely to serve its members effectively. Here, this engagement took the form of systematic, rigorous annotation, discussion, and contextual input from our partner teams in Africa and Western Asia, whose inputs allowed our research team to build effective TFGBV datasets in these languages. In later phases, our technical analysis showed that machine learning classifiers built to detect and prevent TFGBV were more effective when fine-tuned on datasets originating from the target regions and languages.

As we look ahead, we aspire to continue this work with an emphasis on TFGBV in contexts where violent conflict is underway, and to place special attention on Swahili and other sub-Saharan African languages, where there is more room – and dire need — for improvement.

5 A preprint of our academic research paper is available here: https://drive.google.com/file/d/1dP8hjIH_d8GnyMtr-mAylH9E1-56qqCh/