Progress by Social Media Platforms

Google

Background

All four companies (Google, Meta, TikTok, and Twitter) engaged in the Tech Policy Design Labs in 2021 are working on product innovations and prototypes related to OGBV and in line with their commitments. The following changes have been made since TPDL in 2021.

Executive Summary

Google's progress against the commitments is more difficult to assess based on public information, given the variety of entities within the company. Regarding YouTube, we have not seen any announcements suggesting positive steps on curation or reporting in relation to OGBV. Very recently, YouTube announced its YouTube Research Program, providing access to its data and tools to external researchers - which could potentially support OGBV-related research in the future. Jigsaw - a Google entity - developed the Harassment Manager tool in collaboration with Twitter and civil society - the Thomson Reuters Foundation is the first organization to test its use in practice.

Jigsaw – which is a Google entity – developed the Harassment Manager tool in collaboration with Twitter and civil society.

Case study: Google Jigsaw Harassment Manager Tool

Problem addressed:
Women journalists, activists and politicians are facing disproportionate risks of online harassment. 63% of women journalists said they had been threatened or harassed online. Of those, roughly 40% said they avoided reporting certain stories as a result.

Although reporting mechanisms exist in social media platforms, processes and language can make it difficult for victims of abuse to take actions. There was a need for a tool that helps users deal with toxic comments following an incident of harassment, and document their experience.

Approach:
The open source tool Harassment Manager has been developed by Jigsaw (part of Google), as announced in March 2022. This tool aims to help women journalists document and manage abuse targeted at them on social media, starting on Twitter.

More specifically, it helps users identify and document harmful posts, mute or block perpetrators of harassment and hide harassing replies to their own tweets. Individuals can review tweets based on hashtag, username, keyword or date, and leverage a Perspective API to detect comments that are most likely to be toxic.

Stakeholders involved:
This initiative is the fruit of a collaboration between many stakeholders, starting with two tech giants (Google and Twitter). According to Jigsaw, journalists and activists with large Twitter presences have also been involved throughout the whole development cycle. Many NGOs in the journalism and human rights space were also part of this work, including: Article 19, Code for Africa, European Women's Lobby, Feminist Internet, Glitch, International Center for Journalists (ICFJ), Online SOS, Paradigm Initiative, PEN America, Right To Be (formerly Hollaback!), The Thomson Reuters Foundation.

TPDL may have played a role in the development of this tool. Patricia Georgiou, Director of Partnerships and Business Development at Jigsaw, referred to their post-TPLD commitments as an incentive for them: *"Harassment Manager is the result of several years of research, development, and cross-industry collaborations to deliver on our commitment to tackle online violence against women."*

Impact:
The code is now available on Github, open sourced for developers to build and adapt for free. As a first implementation partner, Thomson Reuters Foundation announced in July 2022 the launch of TRFilter, which builds on Harassment Manager's code.