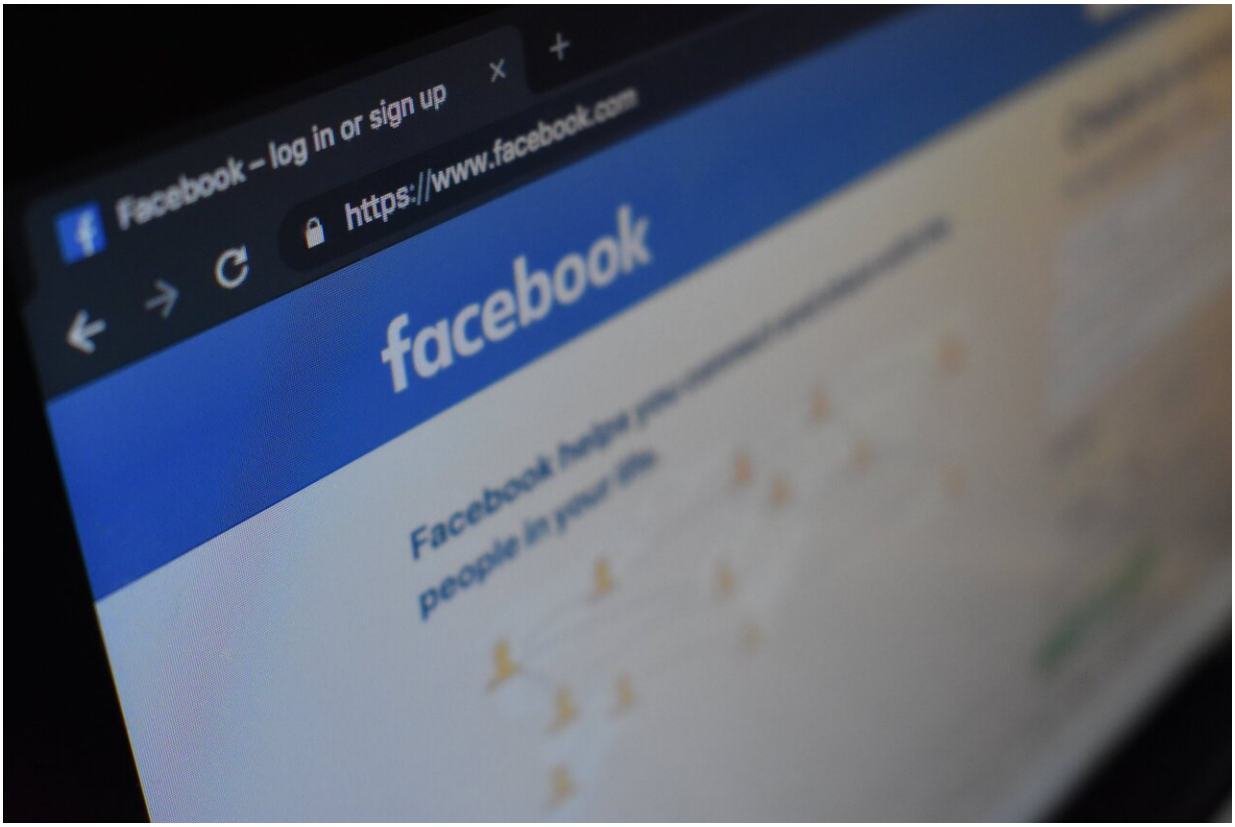
One-size-fits-all content moderation fails the Global South, say researchers

April 13 2023, by Patricia Waldron



Credit: Unsplash/CC0 Public Domain

Social media companies need content moderation systems to keep users safe and prevent the spread of misinformation, but these systems are often based on Western norms, and unfairly penalize users in the Global

South, according to new research at Cornell University.

Farhana Shahid, lead researcher and doctoral student in information science, interviewed people from Bangladesh who had received penalties for violating Facebook's community standards. Users said the content moderation system misinterpreted their posts, removed content that was acceptable in their culture, and operated in ways they felt were unfair, opaque and arbitrary.

"Pick any social media platform and their biggest market will be somewhere in the East," said co-author Aditya Vashistha, assistant professor of information science. "Facebook is profiting immensely from the labor of these users and the content and data they are generating. This is very exploitative in nature, when they are not designing for the users, and at the same time, they're penalizing them and not giving them any explanations of why they are penalized."

Shahid will present their work in April at the Association for Computing Machinery (ACM) CHI Conference on Human Factors in Computing Systems.

Even though Bengali is the sixth most common language worldwide, Shahid and Vashistha found that content moderation algorithms performed poorly on Bengali posts. The moderation system flagged certain swears in Bengali, while the same words were allowed in English. The system also repeatedly missed important context. When one student joked, "Who is willing to burn effigies of the semester?" after final exams, his post was removed because it might incite violence.

Another common complaint was removing posts that were acceptable in the local community, but violated Western values. When a grandmother affectionately called a child with dark skin a "black diamond," the post was flagged for racism, even though Bangladeshis do not share the

American concept of race. In another instance, Facebook deleted a 90,000-member group that provides support during medical emergencies because it shared personal information—phone numbers and blood types in emergency blood donation request posts by group members.

The restrictions imposed by Facebook had real-life consequences. Several users were barred from their accounts—sometimes permanently—resulting in lost photos, messages and online connections. People who relied on Facebook to run their businesses lost income during the restrictions, and some activists were silenced when opponents maliciously and incorrectly reported their posts.

Participants reported feeling "harassed," and frequently did not know which post violated the community guidelines, or why it was offensive. Facebook does employ some local human moderators to remove problematic content, but the arbitrary flagging led many users to assume that moderation was entirely automatic. Several users were embarrassed by the public punishment and angry that they could not appeal, or that their appeal was ignored.

"Obviously, moderation is needed, given the amount of bad content out there, but the effect isn't equally distributed for all users," Shahid said. "We envision a different type of content moderation system that doesn't penalize people, and maybe takes a reformative approach to better educate the citizens on social media platforms."

Instead of a universal set of Western standards, Shahid and Vashistha recommended that social media platforms consult with community representatives to incorporate local values, laws and norms into their moderation systems. They say users also deserve transparency regarding who or what is flagging their posts and more opportunities to appeal the penalties.

"When we're looking at a global platform, we need to examine the global implications," Vashistha said. "If we don't do this, we're doing grave injustice to users whose social and professional lives are dependent on these platforms."

More information: Farhana Shahid et al, Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or WesternPower Hegemony? DOI: 10.1145/3544548.3581538. www.adityavashistha.com/upload … olonial-chi-2023.pdf

Provided by Cornell University