Conflict Prevention on Social Media

Exploring Mechanisms to Mitigate Violence in Online Spaces

Kabui, Charles

2025-05-04

Table of Contents

Figure 1: Social Media Logos

In an era where social media can escalate regional conflicts at unprecedented speeds, platforms are increasingly turning to methods beyond traditional content moderation to prevent violence. As these platforms continue to play a pivotal role in global communication, the urgency for innovative conflict prevention strategies has never been greater. Conventional moderation has often fallen short, particularly in regions with limited platform resources and complex social dynamics.

The Problem: When Social Media Fuels Real-World Violence

Social media platforms have repeatedly been implicated in amplifying regional conflicts, with Meta (formerly Facebook) frequently at the center of criticism. One of the most notable cases occurred in Myanmar, where Facebook acknowledged its role in contributing to the Rohingya genocide [1].

An Independent International Fact-Finding Mission on Myanmar, described Facebook as having played a "*determining role*" in the genocide [2].

[1] In 2018, Meta (formerly Facebook) released a statement accepting the role it played in the genocide against the Rohingya Muslims of Myanmar. "The ethnic violence in Myanmar is horrific and we have been too slow to prevent misinformation and hate on Facebook."

[2] CLASS ACTION COMPLAINT: SUPERIOR COURT OF THE STATE OF CALIFORNIA FOR THE COUNTY OF SAN MATEO

Amnesty International, in its 2022 report "*Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya*" [3], stated that Facebook's algorithms, optimized for engagement, inadvertently prioritized inflammatory content. Compounding the issue, Facebook's limited investment in Burmese-language moderation left harmful content unaddressed during critical periods.

Similar concerns have emerged in Ethiopia [4], India [5], and other regions where linguistic and cultural nuances have transformed seemingly innocuous posts into dangerous triggers.

Current Approaches to Conflict Prevention

Organizations like Data & Society, the Dangerous Speech Project, and various academic institutions have identified several promising approaches now being implemented or studied:

Content Moderation Systems

- **Culturally-aware moderation**: Platforms are investing in region-specific training for both human moderators and AI systems to better understand local contexts.
- **Local language capabilities**: Meta is expanding its support for low-resource languages through initiatives like No Language Left Behind (NLLB) [6].
- **Context-fact notes**: Platforms like Twitter/X flag potentially harmful content based on regional tensions, offering a more nuanced approach [7] [8].

Organizational Approaches

- **Local expertise**: In response to criticism, Meta hired more Burmese-speaking moderators and staff with regional expertise [9].
- **Early warning systems**: Platforms are developing monitoring systems to detect unusual activity patterns that may signal coordinated campaigns or emerging conflicts.
- **Cross-functional crisis teams**: These teams combine engineers, linguists, and conflict resolution experts to respond to crises in real-time.

[3] Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya
[4] What Facebook Does (and Doesn't) Have to Do with Ethiopia's Ethnic Violence
[5] Case Study: Integrity or Influence? Facebook's Governance Trade-offs in India and the Power of the Press
[6] No Language Left Behind (NLLB) is a first-of-its-kind, AI breakthrough project that open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages—including low-resource languages like Asturian, Luganda, Urdu and more.
[7] Updating our approach to misleading information
[8] Meta Turns to Community Notes, Mirroring X
[9] Facebook and Genocide: How Facebook contributed to genocide in Myanmar and why it will not be held accountable

Established Friction Points

Friction points - design elements that deliberately slow user actions to encourage reflection - have shown promise in research:

- **Sharing delays**: Introducing pauses before resharing encourages users to think critically about the content, reducing impulsive misinformation spread [10] [11]. Research has also shown that time pressure can negatively impact the ability to discern true information from false information, while allowing time for deliberation can improve accuracy, according to Nature [12].
- **Read-before-share prompts**: Prompts encouraging users to read articles before sharing have been shown to reduce misinformation spread [13].
- **Content warnings**: A study published in the Journal of Computer-Mediated Communication found that presenting interstitials (brief, neutral messages or images that appears while a chosen website or page is downloading) before or during exposure to inflammatory content can reduce the emotional reactivity to that content [14].
- **Share limits**: WhatsApp's forwarding limits led to a 70% reduction in the spread of highly forwarded messages globally [15].
- **Forwarding friction**: After implementing "*forwarded*" labels and limits, WhatsApp reported a significant decrease in mass message forwarding during India's 2019 elections [16].

---

[10] Pausing to consider why a headline is true or false can help reduce the sharing of false news
[11] Pause before sharing to help stop viral spread of COVID-19 misinformation
[12] Time pressure reduces misinformation discrimination ability but does not alter response bias
[13] How can we combat online misinformation?
[14] Computer-Mediated Communication Preferences and Individual Differences in Neurocognitive Measures of Emotional Attention Capture, Reactivity and Regulation
[15] WhatsApp says viral message forwarding is down 70% after it took steps to combat COVID-19 misinformation
[16] Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-group Ties