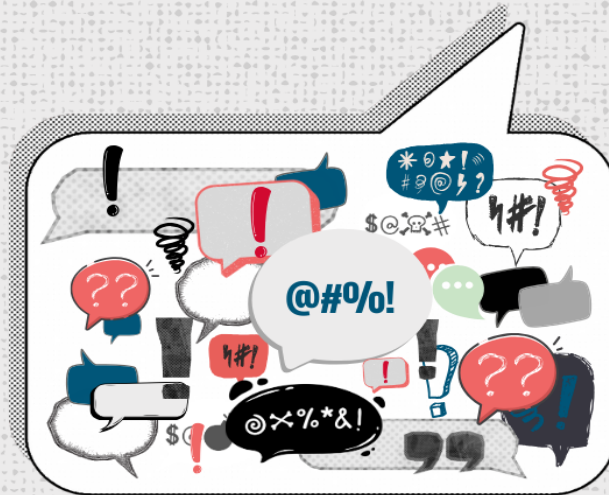




digitally right

Adversarial Narratives Targeting the Rohingya Community in Bangladesh

Table of content

Executive Summary.....	4
1. Introduction.....	5
2. Social media landscape and MDH enforcement.....	7
2.1 Landscape.....	7
2.2 Addressing MDH at platform level.....	8
3. Methodology.....	11
4. What the data tells us about adversarial narratives.....	14
5. Harmful content: What MDH classification reveals.....	18
6. Harmful and dominating narratives.....	22
7. How actors shape narratives.....	32
7.1 Media actors: Fewer posts, outsized reach.....	32
7.2 Political actors: Small footprint, outsized tone.....	35
7.3 Individual actors: High volume, complex dynamics.....	36
8. Platform accountability and policy enforcement.....	38
9. Strategic recommendations.....	42
Reference.....	44
Annex 1.....	47

List of Abbreviations

Abbreviation	Full Form
MDH	Misinformation, Disinformation, and Hate Speech
UNHCR	United Nations High Commissioner for Refugees
ECHO	European Civil Protection and Humanitarian Aid Operations
IOM	International Organization for Migration
APC	Association for Progressive Communications
NGO	Non-Governmental Organization
CSO	Civil Society Organization
AI	Artificial Intelligence
CRI	Centre for Research and Information
BTRC	Bangladesh Telecommunication Regulatory Commission
EFF	Electronic Frontier Foundation
RTV	Radio Television Vision (Bangladesh-based media)
BBC	British Broadcasting Corporation
GDI	Global Disinformation Index
Meta	Parent company of Facebook and Instagram
BJP	Bharatiya Janata Party (India)
AAP	Aam Aadmi Party (India)
BNP	Bangladesh Nationalist Party

Executive Summary

This report investigates how online narratives about the Rohingya have evolved on social media platforms in Bangladesh over a three-month period from December 2024 to February 2025. The study, based on over 6,300 posts collected mainly from Facebook and YouTube, shows that about one in every five posts carried harmful, adversarial narratives aimed at ridiculing, dehumanizing, or inciting distrust toward the Rohingya community. Facebook proved to be the primary driver of such hate speech, hosting the vast majority of the harmful content. Although YouTube contributed a smaller portion of posts, it was especially effective in amplifying high-severity disinformation, garnering significant views for videos that presented false claims.

The report goes to show that narratives targeting the Rohingya typically use emotionally charged language that frames them as national security threats or as agents of destabilization. These narratives are reinforced by various actor types. While individual users produce the majority of harmful posts, media actors, though few in number, generate substantially higher engagement, and political actors, despite their limited output, see their messages spread widely. In combination, these patterns reveal that harmful content often achieves high visibility and rapidly influences public opinion, despite its relatively low volume.

The report also highlights critical gaps in platform accountability. Both Facebook and YouTube have policies intended to curb hate speech and misinformation, yet the report finds that harmful content written in Bangla remains largely unmoderated. This lapse is heightened by challenges in automated detection systems and limited regional language support, suggesting that current moderation tools are inadequate for addressing the unique ways in which hate speech and disinformation spread locally.

Based on these findings, the report calls for strategic actions that include investing in refugee-sensitive journalism, enhancing media literacy programs beyond traditional classrooms, and developing faster, more reliable channels for content moderation. In addition, it emphasizes the need for cross-border cooperation and transparency in algorithmic moderation to counter the spread of harmful narratives effectively. Ultimately, the report stresses that a human rights-centered and community-driven approach is essential to protect the Rohingya community from the corrosive effects of digital hate.

This report is part of a broader pilot initiative led by the United Nations High Commissioner for Refugees (UNHCR), with support from the European Commission's humanitarian arm (ECHO), to strengthen protection responses against digital harm.

1. Introduction

Bangladesh is home to more than one million Rohingya refugees who fled persecution in neighboring Myanmar ([UNHCR, n.d.](#)). Most now live in 33 overcrowded camps in the southeastern district of Cox's Bazar ([IOM, 2024](#)), with only a small number relocated to a remote, government-built facility on Bhashan Char, an island in the Bay of Bengal. Lacking legal status, their movement outside the camps is restricted, leaving them almost entirely dependent on humanitarian aid ([UNHCR, n.d.-a](#)).

Rohingya migration into Bangladesh began decades ago, rooted in long-standing discrimination by successive governments in Myanmar ([CFR, 2020](#)). The largest wave came in 2017, after a brutal military crackdown in Myanmar's Rakhine State forced more than 700,000 people to cross the border into Bangladesh, seeking safety from violence and statelessness ([UNHCR, 2017](#)).

Though lives were saved, the situation for the million Rohingya refugees in Bangladesh remains precarious. Humanitarian aid and government support have helped address urgent and basic needs in the camps, but conditions have steadily declined. Limited resources, environmental risks, rising insecurity, and growing tensions with local communities have deepened the challenges, leaving many unsure of what lies ahead. ([Human Rights Watch, 2021](#)). An ongoing war in Myanmar only exacerbated the uncertainty.

Public perception in Bangladesh has also shifted over time. In the immediate aftermath of the 2017 crisis, the government and the public alike showed strong solidarity with the Rohingya. Moved by what many saw as anti-Muslim violence across the border, Bangladeshis offered food, medicine, and shelter to those fleeing Myanmar. But as the years have passed, attitudes have hardened against them ([Ansar & Khaled, 2021](#)). What began as a humanitarian response has turned into growing resentment.

Politicians and segments of the media now portray the refugees as drug traffickers or potential security threats. The government, once vocal in its support, increasingly frames the crisis as a strain on the economy and a risk to political stability ([France, 2022](#)), a discourse shaped by rising economic pressures and concerns over security in areas surrounding the camps ([The Daily Star, 2023](#)).

Tensions rooted in the material world often find more volatile expression online—amplified, distorted, and weaponized. In recent years, a more corrosive force has shaped public sentiment: the spread of online misinformation and hate speech. False claims, emotionally charged narratives, and disinformation have been used to manipulate perception, justify discriminatory policies, and entrench hostility toward the Rohingya. They have been reinforced through a network of government media, social platforms, religious rhetoric, and even popular literature, contributing to the community's continued marginalization ([Khatun, 2024](#)).

Looking back at the 2017 crackdown on the Rohingya, the role of social media platforms in amplifying hate and justifying violence through disinformation is well documented. In 2018, U.N. human rights chief Marzuki Darusman accused platforms like Facebook of playing a key role in the genocide carried out against the Rohingya in Myanmar ([Reuters, 2018](#)). Before 2016–17, Facebook was virtually the only social media platform widely used in Myanmar. It was during

this period that the military launched so-called “clearance operations” targeting Rohingya Muslims ([Tähtinen, 2024](#)).

The U.N. Human Rights Council later accused the Myanmar military of war crimes and crimes against humanity, citing its disproportionate and unlawful attacks on ethnic and religious minorities. A U.N. Fact-Finding Mission concluded that the situation met the legal threshold for genocide and called for prosecution of senior military officials ([Schissler, 2024](#)).

While the 2017 atrocities in Myanmar are well documented, far less attention has been paid to how misinformation, disinformation, and hate speech (MDH) have shaped public opinion in Bangladesh toward the Rohingyas. Despite growing tensions in host communities, research into how social media platforms have fueled hostility remains limited.

This report seeks to fill part of that gap, within a limited scope. It focuses on Facebook and YouTube—the two most widely used platforms in Bangladesh—as primary channels through which harmful narratives against Rohingya refugees circulate. The report maps the key actors, messages, and patterns driving rising hostility, and examines the broader implications for social cohesion and human rights. Its findings are based on three months of regular monitoring of posts containing hate speech and disinformation.

The report, supported by the Association for Progressive Communications (APC), aims to aid the United Nations High Commissioner for Refugees (UNHCR) and its partners to respond to online harms targeting the Rohingya refugees.

2. Social media landscape and MDH enforcement

2.1 Landscape

To understand how online hate speech and disinformation gain traction, it is essential to consider the broader social media landscape in Bangladesh. According to the Bangladesh Telecommunication Regulatory Commission, more than 130 million people now have internet access—roughly three out of every four residents, by the government’s definition ([BTRC, n.d.](#)). While mobile internet growth has slowed in recent months, due in part to rising taxes, broadband subscriptions continue to climb ([The Daily Star, 2023b](#)).

Social media use has expanded alongside this digital growth. As of January 2025, there were 60 million active social media user identities in Bangladesh—a 13% increase over the previous year, according to Data Reportal, which compiles data from platform advertising systems ([Kemp, 2025](#)).

Among platforms, Facebook remains dominant, with 60 million users—about one-third of the country’s population ([Kemp, 2025](#)). YouTube follows closely with 44.6 million users, while TikTok has recently overtaken YouTube, reaching 46.5 million users aged 18 and above. While the figures do not reflect unique individuals—since a single person can operate multiple accounts or use different devices—they offer a reliable estimate of how many user identities are being reached through social media advertising.

Table 1: Social media users in Bangladesh

Facebook	Tiktok	YouTube	Instagram
60 million	46.5 million	44.6 million	7.5 million
			

Source: Digital 2025: Bangladesh, Data Reptotal

In Bangladesh, social media platforms have emerged as primary sources of news and information—but also as significant channels for spreading disinformation and hate. This dynamic has allowed various political and commercial actors to shape public narratives strategically. In May 2024, Meta, Facebook's parent company, removed 50 accounts and 98 pages originating in Bangladesh, linked to individuals associated with the ruling Awami League party and the Centre for Research and Information (CRI), a nonprofit organization, for engaging in coordinated inauthentic behavior ([The Daily Star, 2023c](#)).

Bangladesh was also among the top ten countries globally in the final quarter of 2024 for YouTube content removals. YouTube took down 272,216 videos originating from the country due to violations of Community Guidelines, which cover scams, fraud, explicit content, harassment, disinformation, hate speech, and incitement to violence. While YouTube does not provide detailed country-specific breakdowns for the types of content removed, its global statistics indicate about 1 percent of channels and 2.5 percent of videos were removed specifically for hateful content ([Google Transparency Report, n.d.](#)). Meanwhile, Digitally Right's information research unit, Dismislab, documented multiple cases of financially motivated actors in Bangladesh spreading disinformation on YouTube, with the platform often indirectly profiting from such activities ([Misinformation on YouTube: High Profits, Low Moderation, 2024](#)).

2.2 Addressing MDH at platform level

Both Meta and Youtube have policies to combat MDH. [Meta's Community Standards](#) explicitly prohibit harmful content and hate speech (Meta, n.d.-a). Facebook employs a combination of artificial intelligence and human moderators to identify and remove such content. It collaborates with third-party fact-checkers to label misinformation and reduce their spread.

[YouTube's Community Guidelines](#) prohibit hate speech, defined as content promoting violence or hatred against individuals or groups based on attributes such as race, ethnicity, or religion (YouTube, n.d.). It also has policies against misinformation, particularly content that poses serious risks of egregious harm. It uses automated systems and human reviewers to detect and remove violative content.

However, Facebook's automated content moderation has shown significant gaps in non-English languages, including Bangla ([Heath, 2021](#)). YouTube also faces challenges, with studies highlighting the need for improved hate speech detection methods in Bengali ([Faria et al., 2024](#)).

Both platforms have also been criticized for not doing enough to curb the spread of harmful content related to the Rohingya crisis ([Amnesty International, 2023](#)). Critics point out that YouTube videos inciting violence or spreading false information about the Rohingya remained accessible on the platform, potentially exacerbating tensions ([Chimkhilai & Panyametheekul, 2024](#)).

Table 2: Policy and enforcement comparison: Facebook vs YouTube

Category	Facebook (Meta)	YouTube (Google)
Hate Speech Definition	Direct attacks against people on the basis of protected characteristics including race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.	Promoting violence or hatred against individuals or groups based on protected attributes such as race, ethnicity, religion, nationality, immigration status, gender identity, or sexual orientation.
Prohibited Content Types	- Dehumanizing speech (e.g., animal comparisons) - Statements of inferiority - Calls for exclusion/segregation	- Inciting hatred or violence - Dehumanizing language - Slurs and stereotypes targeting protected groups
Enforcement Mechanisms	- Automated detection (AI/ML) - Human moderation - User reports - Global and local fact-checking partnerships (Except US)	- Machine learning detection - Human reviewers - Community flagging - Context-specific policy enforcement
Moderation Actions	- Remove violating content - Label or reduce reach of borderline content - Disable accounts or pages for repeated offenses	- Remove content - Age-restrict or demonetize videos - Reduce recommendations for borderline content
Misinformation Policy	Targets content causing physical harm (e.g., anti-vaccine, civic disruption) Labels false content via third-party fact-checkers (Meta)	Removes content that could cause “egregious harm,” such as incitement or medical misinformation (YouTube)

Meta’s own transparency data shows a clear and steady decline in the amount of hate speech content the company has taken action on since its peak in 2021. From a high of over 30 million removals, the number has dropped sharply to just under 5 million by early 2025. (*Meta Transparency Center – Hate Speech Content Actioned, 2025*) This downward trend raises serious concerns, especially in light of recent policy changes that further weaken content moderation standards.

The recent policy shifts at Meta mark a troubling step back in the company’s approach to content moderation—one that has implications for vulnerable communities ([Officer & Meta, 2025](#)). The rollback of key safeguards, such as restrictions on hate speech and the use of automated moderation tools, risks creating an environment where harmful content can spread more freely. Amnesty International has warned that these changes “drastically increase the risk that the company will yet again contribute to mass violence and gross human rights

abuses—just like it did in Myanmar in 2017” ([Amnesty International. 2025](#)). They argued that Meta’s algorithms already prioritize content that triggers emotional reactions—often at the expense of human rights and safety—and that further loosening of rules will only heighten the spread of hate and disinformation targeting their community ([Amnesty International. 2025b](#)).

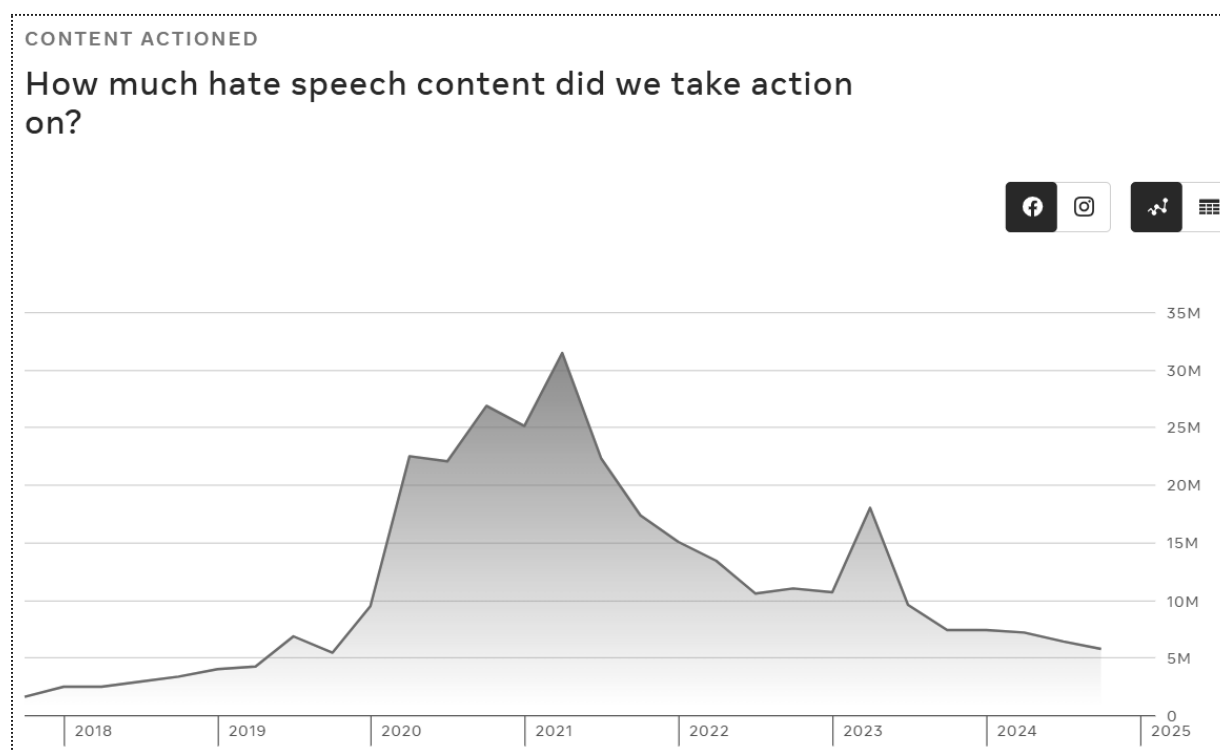


Chart 1: Meta’s enforcement of hate speech content has declined sharply since its peak in 2021. Source: Meta Transparency Center (Accessed April 2025) ([Meta, n.d.-b](#)).

For countries like Bangladesh, where moderation in local languages such as Bangla or Rohingya is already limited, the retreat from automation without a corresponding investment in human moderation presents an even greater risk. According to the Electronic Frontier Foundation (EFF), if Meta is pulling back its use of automation, it must “invest globally in qualified content moderation personnel to make sure that content from countries outside of the United States and in languages other than English is fairly moderated” ([Greene. 2025](#)).

3. Methodology

This section explains how the research was carried out and describes how content was collected from Facebook and YouTube, how it was organized and categorized, and how the team analyzed adversarial and MDH narratives about the Rohingya community in Bangladesh.

Monitoring

To understand how online discourse is shaping public sentiment toward the Rohingya community, this study focused on two of Bangladesh's most widely used social media platforms: Facebook and YouTube. Over a three-month period—from December 1, 2024, to February 28, 2025—a team of two trained research assistants conducted daily monitoring of posts containing keywords linked to the Rohingya.

For YouTube, an automated content scraper was used, allowing for large-scale data collection. Facebook, however, posed a greater challenge. Due to the platform's technical restrictions and limited accessibility for researchers, data had to be collected manually. Researchers identified, reviewed, and recorded the posts, one by one based on relevance and the presence of predefined keywords in the search results.

Building the dataset

The data collection process was anchored in a curated list of Bangla and Bengali-language keywords, many of which were spelled in the English alphabet—a common practice in local social media discourse. Variations included different Bangla and Bang-lish renderings of the word "Rohingya," such as "রোহিঙ্গা," "রোহিংগা," and "রোয়াইংগা." Posts containing any of these keywords were systematically stored in structured spreadsheets.

YouTube data, collected automatically, was later compiled and cleaned for analysis. Facebook content, gathered from the search tool Who Posted What, was categorized and logged in a similar format. All data was anonymized and securely stored to ensure compliance with ethical research standards.

Categorization

The research applied a dual-framework approach to make sense of the wide range of content collected. The first, adapted from the Global Disinformation Index (GDI), focused on identifying adversarial narratives in three categories: 1) negative targeting, 2) sensationalization, and 3) in-group versus out-group framing. GDI's approach allows for the analysis of content beyond rigid definitions of misinformation or hate speech, offering a narrative-based lens to understand how certain messages influence public attitudes.

To better analyze the structure and intent of adversarial narratives, the study further organized the content into eight subcategories under the three main types. In-group/out-group framing included posts that criminalized the Rohingya, called for harm, or portrayed them as outsiders through stereotyping and discrimination. Negative targeting captured ridiculing language, slurs,

and content that spread distrust. Sensational content used fearmongering and sweeping or generalized claims to stir strong emotions.

The adversarial posts were also categorized based on the Misinformation, Disinformation, and Hate Speech (MDH) framework, to distinguish the different forms through which harmful information is spread.

- Misinformation included false, misleading, or exaggerated claims intended to sow distrust or reinforce harmful stereotypes.
- Disinformation: High severity misinformation with wider scope of harm and malicious intent were categorized as disinformation
- Hate speech refers to explicit or implicit messages that promote hostility, exclusion, or harm toward the Rohingya community.

Review

The research combined quantitative metrics with qualitative insights. Quantitative analysis tracked how often different types of harmful content appeared and how those frequencies changed over time. Meanwhile, narrative analysis dove deeper—examining the evolution of themes, the context in which harmful narratives emerged, and the identities or affiliations of key actors driving the conversation.

Each content was rated using a three-tier severity scale. Posts with misleading or exaggerated claims were classified as misinformation at low or moderate levels, while high-severity false claims that displayed intent to harm were categorized as disinformation.

Hate speech was assessed in the same way. At the lower end were posts with discriminatory or offensive language. More severe examples included dehumanizing or hostile language, and the most serious involved content that incited violence or promoted harm.

Adversarial content that had already been debunked by professional fact-checkers or verified by credible media sources were considered as misinformation. In most cases claims were independently fact-checked by the reviewers with experience in online verification.

Three independent researchers were tasked with reviewing and categorizing the data to address potential bias. In cases of disagreement, reviewers discussed their findings and reached consensus. Some content was subjected to blind review, with the reviewers unaware of the post's original source, to avoid subconscious bias.

Limitations

While the study provides a useful look into how harmful narratives about the Rohingya appear on social media, it comes with a few important limitations.

First, the data collection relied on specific keywords in Bangla and Banglish. This helped focus the monitoring but may have missed content that used different spellings, slang, or coded language. Subtle or indirect narratives—especially those that evolve over time—may not have been captured.

The tools used also had limits. On YouTube, data was collected automatically, but the volume was relatively small—around 250 to 350 posts per month—compared to the vast amount of content online. These tools were designed more for analyzing themes and tone than for measuring scale, which limited the depth of quantitative analysis.

Facebook presented additional challenges. Because automated scraping is restricted, all data had to be collected through manual search on Who Posted What. Such searches only show public posts, in results. Private groups and personal profiles remain excluded—spaces where sensitive or harmful content is often shared. Therefore, some key narratives likely went unseen.

Although reviewers followed a shared framework, content analysis is never fully objective. Judgments about tone, context, or intent can vary, even among experienced researchers. Efforts were made to reduce bias, but some inconsistencies are always possible.

These limits should be considered when reading the findings. They point to the challenges of tracking digital harm in real time—and the need for continued investment in more robust and flexible monitoring tools.

4. What the data tells us about adversarial narratives

Monitoring data from Facebook and YouTube offers a bird's-eye view of how narratives around the Rohingya community spread and evolve online. It helps translate scattered digital discourse into a clearer picture of trends, tone, and influence.

Between December 2024 and February 2025, the study tracked 6,356 public posts across Facebook and YouTube. Of these, Facebook accounted for 71%, underscoring its continued dominance in Bangladesh's information ecosystem. YouTube contributed 29% —a smaller but still significant portion, particularly given the higher visibility and engagement with video content.

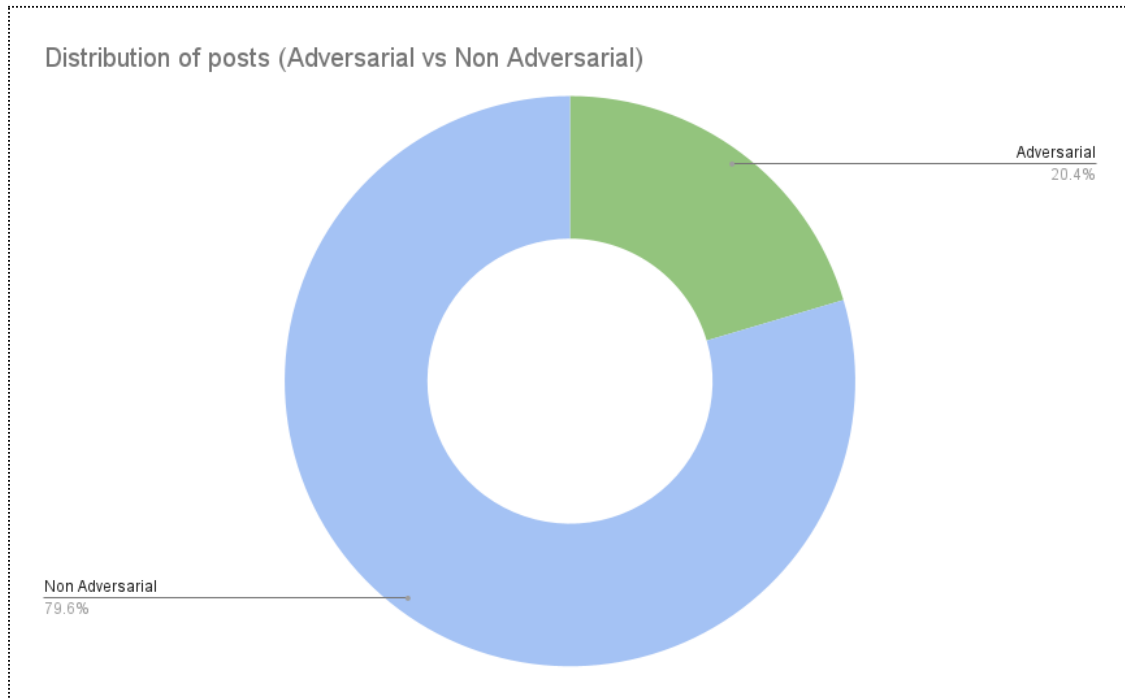


Chart 2: Out of 6,356 posts analyzed, 20.4% were adversarial while 79.6% were non-adversarial, indicating that harmful narratives made up roughly one-fifth of the total content.

One in five posts targeting Rohingya refugees carried adversarial narratives, with Facebook dominating volume and YouTube driving reach

Among all posts, 1,297 (20%) were categorized as adversarial. This means that one in every five posts about the community actively sought to ridicule, dehumanize, or incite distrust. The vast majority (94%) coming from Facebook and only 6% from YouTube.

Nearly half (45.8%) of all adversarial posts featured negative targeting narratives—content that included ridicule, slurs, or messages designed to erode trust. Another 36.5% had in-group and out-group framing, portraying the Rohingya as outsiders and putting them in opposition to the “in-group”—the host society. Sensational content made up the remaining 17.7%, typically using emotionally charged language and sweeping generalizations to stir reactions.

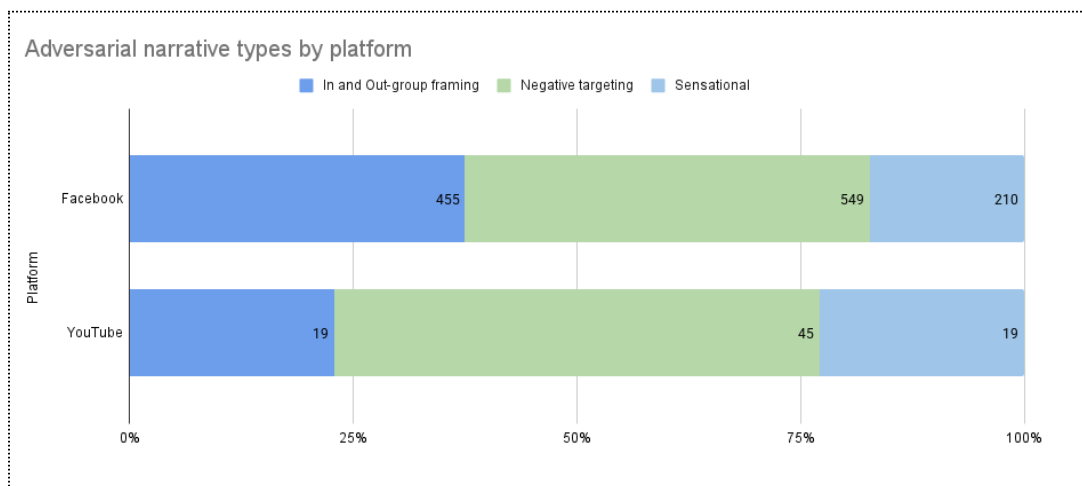


Chart 3: Adversarial narratives on Facebook are more varied, with high counts across all three categories. YouTube content is more concentrated in negative targeting, with limited use of in-group framing and sensationalism.

Among the 1,297 adversarial posts, nearly one in four featured “Ridiculing” narratives (25%), followed closely by “Othering” (24%). These posts often mocked Rohingya individuals or portrayed them as fundamentally different, reinforcing exclusion. “Seeding distrust” appeared in 14% of posts, while “Sweeping generalizations” made up 11%, and “Criminalizing” narratives—linking Rohingyas to crime or insecurity—accounted for 6%.

Some of the most damaging content wasn’t the most common—but the most visible

When examining which types of narratives captured the most attention, the picture became more complex. Narratives that seeded distrust, made sweeping generalizations, or criminalized Rohingyas consistently drew higher engagement across platforms—despite appearing less frequently. For example, sweeping generalizations accounted for only 12% of posts, but averaged 6,450 views on YouTube and 127 reactions on Facebook. Distrust-seeding content, which made up 15% of posts, received the highest average reactions on Facebook (249) and over 1,928 average views on YouTube.

Criminalizing narratives accounted for just 6% of the total volume, but averaged 3,285 views on YouTube and 159 reactions on Facebook—consistently placing them among the most engaging subcategories.

In contrast, high-volume categories like Ridiculing and Othering—which made up 25% and 24% of adversarial posts respectively—garnered less attention. Ridiculing content averaged 99 reactions on Facebook and 5,877 views on YouTube, while Othering posts averaged just 65 reactions and 85 views.

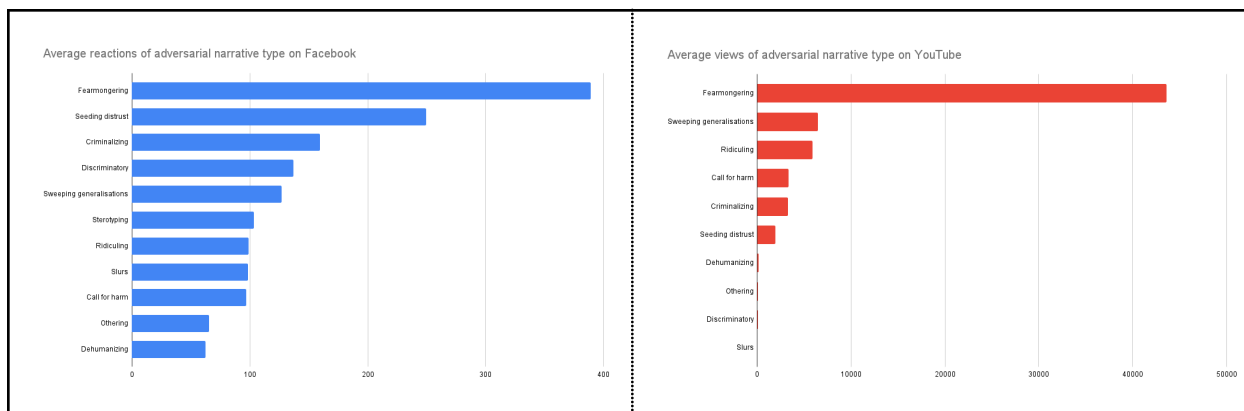


Chart 4: Average audience engagement with adversarial narrative types on Facebook and YouTube. The left chart displays the average number of reactions per post for different adversarial narrative subcategories on Facebook, while the right chart shows the average number of views per video for the same categories on YouTube. On both platforms, fearmongering narratives receive the highest levels of audience engagement.

One post categorized as “call for harm” on YouTube received over 3,000 views—making it the top-performing post in that subcategory on the platform, though it represents a single post only. The same type of posts received lower reactions on Facebook, averaging only 97.

Media actors, though fewer in number, command disproportionate influence through higher engagement per post.

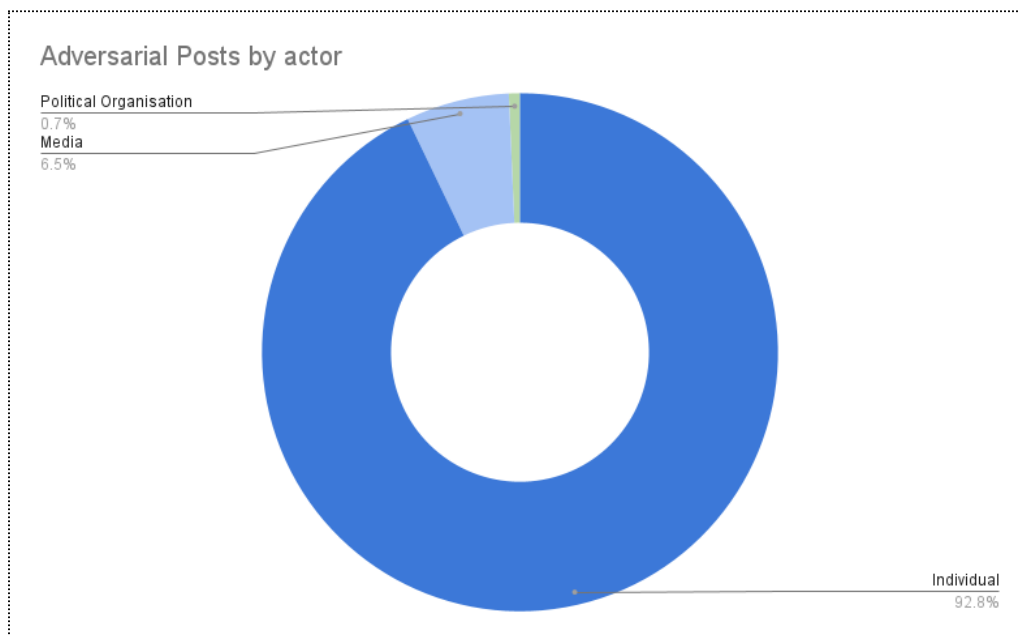


Chart 4: The vast majority of adversarial posts—92.8%—were shared by individual accounts, with media outlets responsible for 6.5% and political organizations for just 0.7%.

Although most content came from individuals, media actors were more effective in drawing reactions and views—especially when using highly emotive or sensationalized framing. Individual accounts produced 93% of all adversarial posts, while media organizations contributed 6.6% and political organizations only 0.6%. Despite the smaller footprint, media actors generated significantly higher average engagement. On Facebook, individual actors averaged 87 reactions per post, while media pages saw 865. On YouTube, media posts averaged 7125 views per video—more than five times that of individual actors, who averaged 1,201 views.

Different actors favored different types of narratives. Individuals most frequently used ridicule (26%), othering (25.2%), and seeding distrust (12.2%). Media actors leaned toward sweeping generalizations. Political organizations, though few in number, focused on othering and ridiculing content mostly.

Table 3. Top narrative subcategories used by each actor type	
Actor Type	Top Narrative Subcategories
Individual	Ridiculing, Othering, Seeding distrust
Media	Seeding distrust, Sweeping generalisations
Political	Othering, Ridiculing

The media category also influenced narrative patterns. National media, for example, posted less (24.1% of all media posts) but received higher average engagement per post—over 1,100 reactions on Facebook and 8,600 views on YouTube. These accounts leaned toward negative targeting (50%), while local media favored sensationalism (37%).

5. Harmful content: What MDH classification reveals

The MDH framework—short for Misinformation, Disinformation, and Hate Speech—helps clarify how online content shapes public attitudes toward the Rohingya community. While all three content types can overlap in impact, each functions differently. Misinformation often spreads factually incorrect information without clear intent, disinformation reflects deliberate manipulation, and hate speech promotes hostility or exclusion. Reviewing content through this lens sheds light on how different kinds of digital harm circulate—and what impact they have.

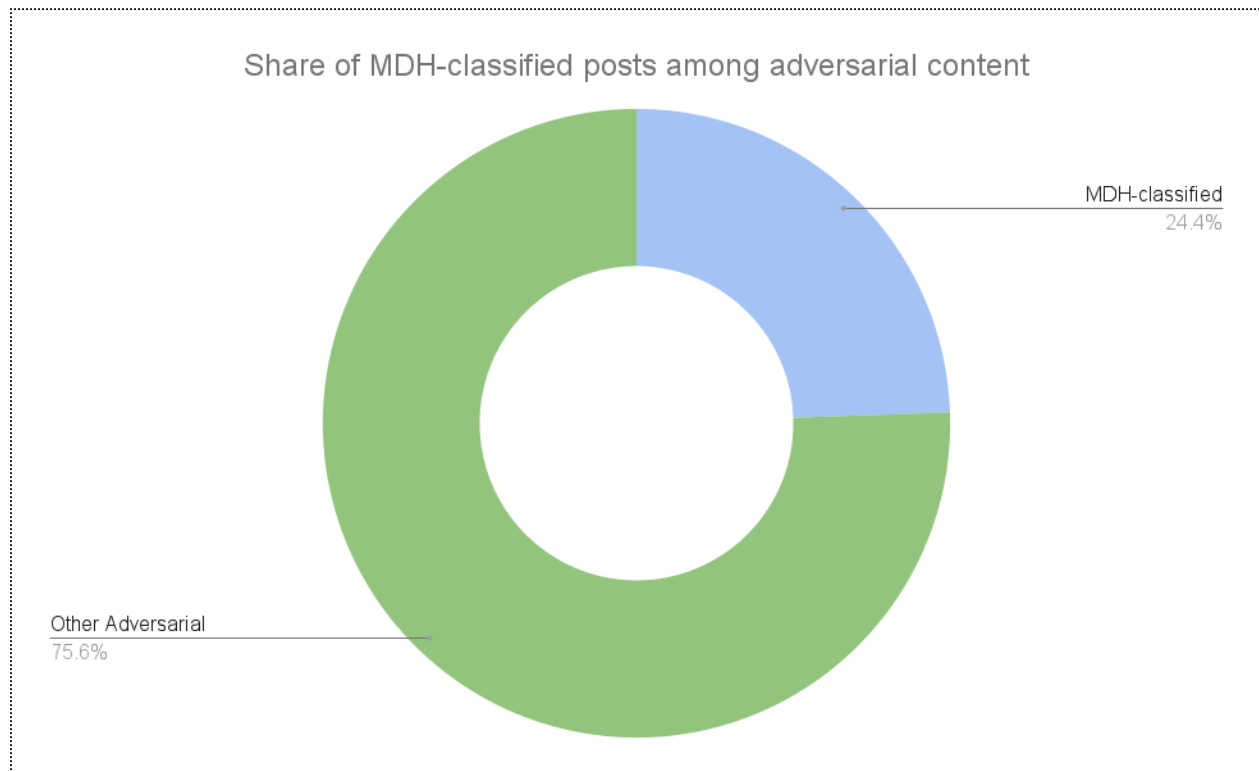


Chart 6: Out of 1,297 adversarial posts in the dataset, just 24.4% were categorized under the MDH framework

Out of 1,297 adversarial posts in the dataset, 24.4% (317 posts) contained clear elements of misinformation, disinformation, or hate speech. While all posts were adversarial, most were not classified using this stricter content-type lens.

Hate speech remains the dominant threat, especially on Facebook

Of the adversarial content analyzed, hate speech was the most prevalent type by a wide margin. Nearly 71% of all MDH-classified posts were tagged as hate speech, far outpacing both

misinformation and disinformation. This trend was especially visible on Facebook, where 203 out of 286 tagged posts contained hate speech. These included slurs, dehumanizing language, and explicit or coded calls for exclusion.

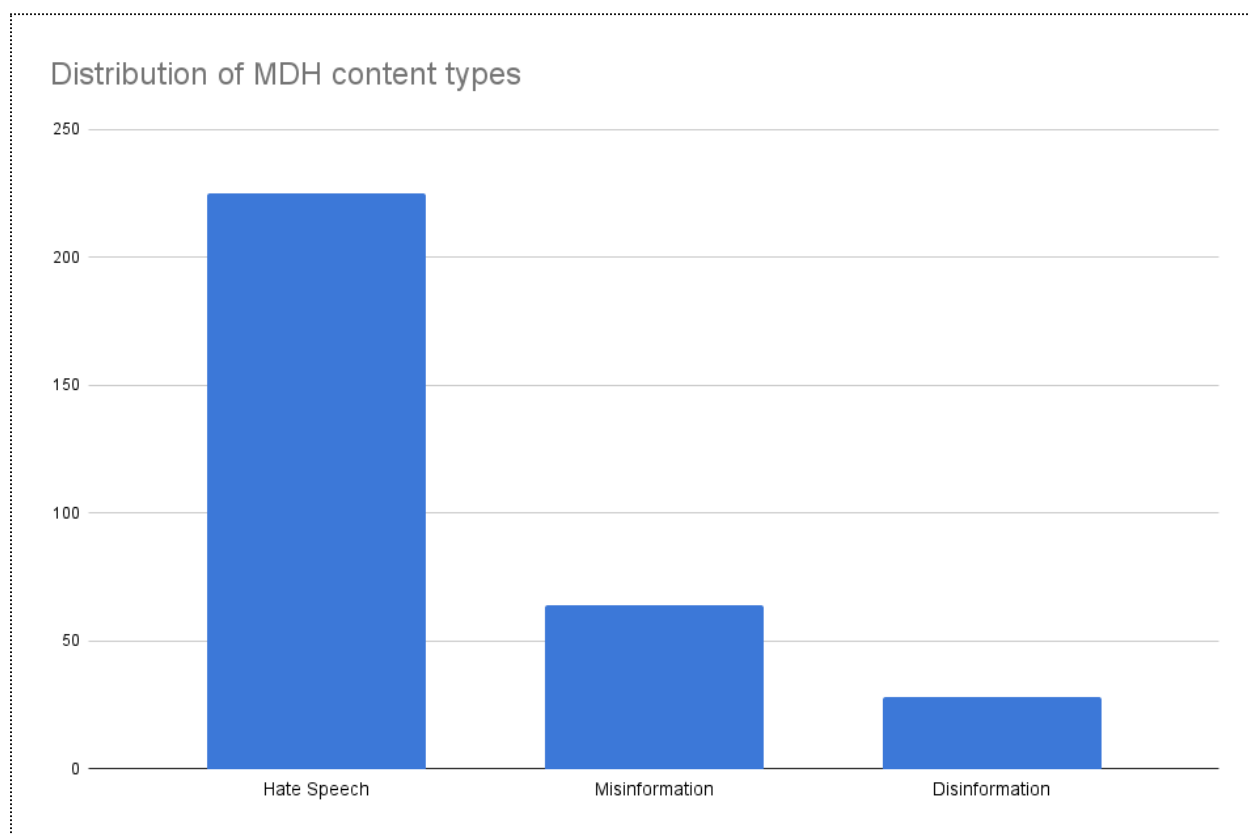


Chart 7: Hate speech made up nearly 71% of all MDH-classified posts, with far fewer tagged as misinformation or disinformation

The intensity of engagement further underlined the spread of this type of content. On Facebook, hate speech posts averaged over 176 reactions each. Many of these posts framed the Rohingya in derogatory terms or falsely linked them to disease, crime, or social disruption. On the other hand, the volume of misinformation posts on Facebook was relatively low, but generated a higher level of engagement. On average, each misinformation post received 341 reactions on the platform.

YouTube videos tagged as disinformation attract the highest viewership

While Facebook hosted the highest volume of MDH content, YouTube posts tagged as disinformation drew the most viewers. On average, each disinformation video received 128632 views—making them the most-watched content type on the platform. These videos often promoted conspiracy theories or claims that had been publicly debunked, including allegations of terrorist training in refugee camps or foreign interference linked to the Rohingya presence.

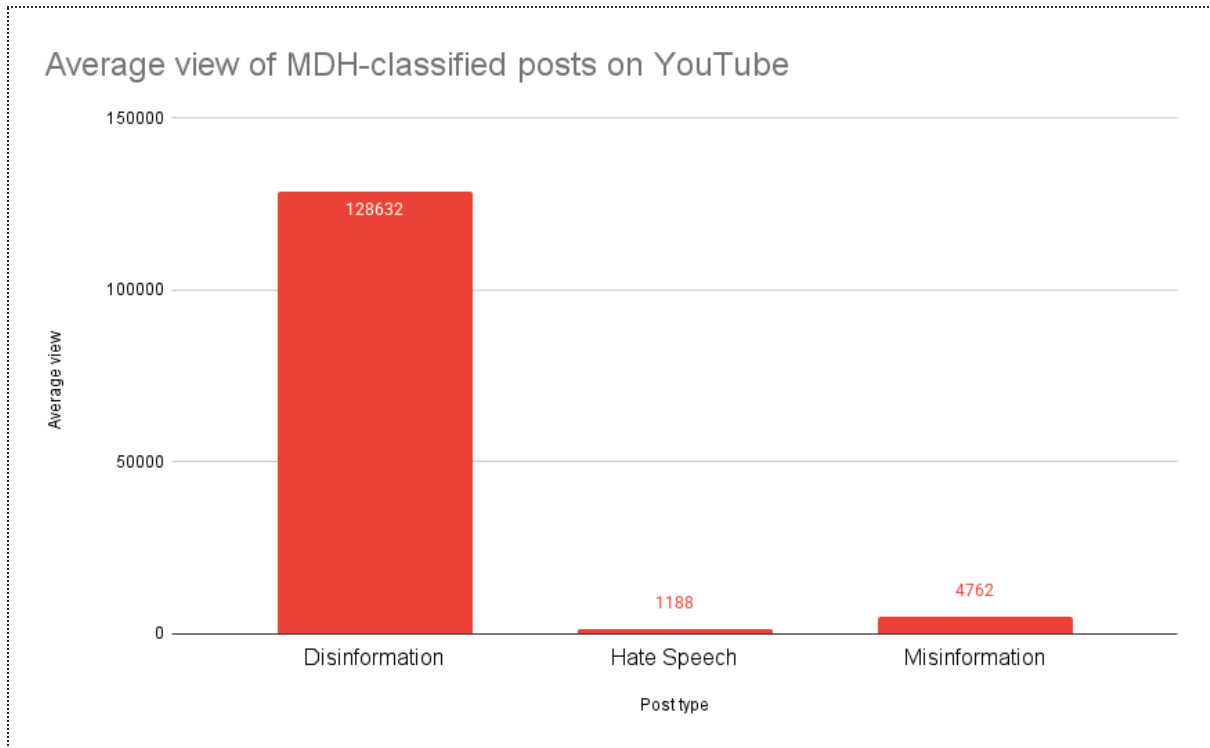


Chart 8: Disinformation content on YouTube drew the most views per video, despite its lower volume. On average, each disinformation video received 128,632 views.

Individual actors produce the most MDH content—but media pages drive higher per-post engagement

Despite making up a relatively small share of total posts, disinformation content leveraged YouTube’s algorithmic visibility and longer video format to spread misleading narratives. Some actors embedded these claims in documentary-style presentations, making them appear more authoritative.

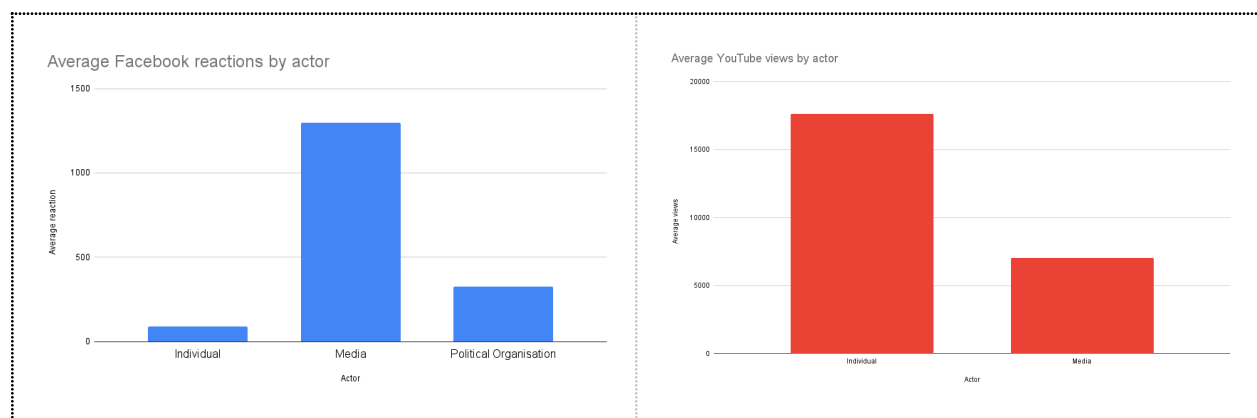


Chart 9: While individuals contributed most posts, media actors consistently drove more engagement per post. On both Facebook and YouTube, media outlets had a disproportionate influence on visibility—even with fewer uploads.

Individual users were responsible for nearly 88% of all MDH-classified posts. This reflects broader patterns across the dataset, in which everyday users were the primary source of harmful narratives. These posts often blended misinformation and hate speech in ways that blurred intent—sometimes repeating rumors without clear awareness of their falsity, and at other times deploying overtly hostile rhetoric.

But despite their numerical dominance, individual actors trailed far behind media outlets in engagement per post. Media pages, while accounting for only 36 posts in total, generated far more reactions and views on average. On Facebook, hate speech posted by media outlets received more than triple the engagement of similar posts by individuals. This suggests that harmful narratives, when amplified by media accounts, can gain traction faster—even if they appear less frequently.

High-severity content drives disproportionate engagement—especially on YouTube, and particularly when amplified by media actors



Chart 10: Posts rated as more severe received more engagement, especially on YouTube. High-severity videos often include fear-based or dehumanizing narratives, and their viral potential highlights platform-level risk.

Posts with the most extreme language didn't always get the most attention—though when they did, the impact was clear. On YouTube, a single Level 3 video averaged over 3,300 views, making it the most viewed by severity. In contrast, Level 1 videos—more common but less intense—averaged around 1,580 views, while Level 2 videos lagged at just 278 views.

On Facebook, the pattern flipped: Level 1 posts drew the highest average reactions (229), compared to 147 for Level 3 and only 60 for Level 2, implying that on Facebook, milder posts may spread further, while on YouTube, sharper content—though rare—lands harder when it goes viral.

The data suggests that whether through algorithmic prioritization or social sharing patterns, emotionally extreme content appears more likely to spread, raising questions about how moderation systems respond to the harmful content.

6. Harmful and dominating narratives

This chapter explores the nine recurring and particularly harmful narratives surrounding the Rohingya during the three-month monitoring period, focusing on misinformation, disinformation, and hate speech (MDH). While numerous narratives circulated, we selected these nine because of their frequency and the significant harm they caused. It analyzes how MDH, along with broader hostile narratives, worked to frame the Rohingya community as outsiders, threats, or scapegoats.

Though the original scope of the study was focused on content within Bangladesh, the findings revealed a broader trend. Many of the most damaging narratives, especially those linked to political manipulation or fears of demographic change, originated from the Bangla-speaking regions of India. Posts from politically affiliated individuals and fringe media outlets there garnered some of the highest engagement levels across the dataset.

These cross-border dynamics show how anti-Rohingya narratives transcend national boundaries, traveling through a shared language and digital ecosystem. From calls for violence to the subtle weaponization of everyday language, the following sections delve into how identity, fear, and misinformation are used to normalize exclusion and, in some cases, justify harm.

To give a fuller picture, we've also included some positive perspectives and posts expressing support for the Rohingya, offering a counterpoint to the harmful narratives discussed.

a) Infiltration narratives frame Rohingyas as a threat to national identity to stoke fear-based population panic



Between December 2024 and February 2025, a wave of posts across Facebook and YouTube stoked fears of mass infiltration by the Rohingya population—not only into Bangladesh but also into India. Posts claimed imaginary waves of “millions” or even “crores” of undocumented Rohingya refugees crossing borders, often portraying them as destabilizing agents or pawns of political parties.

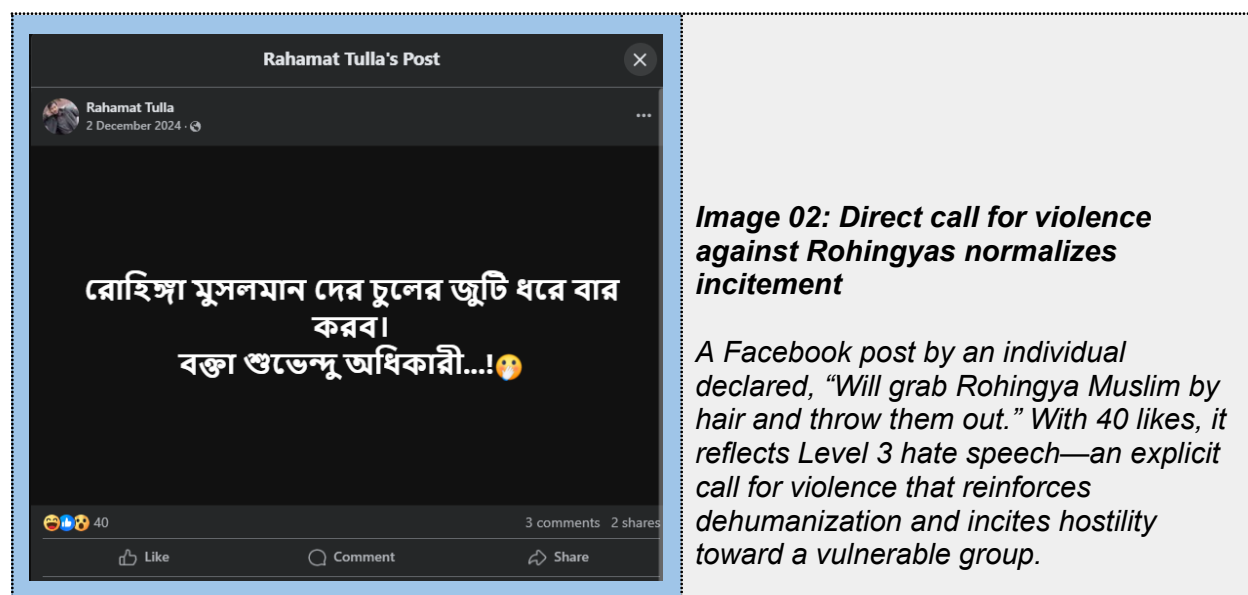
One Facebook post from December 2024 claimed, “Crores of Rohingyas have infiltrated West Bengal, Delhi, and Hyderabad, changing the demographic fabric.” The post further claimed that political parties were using them as vote banks, suggesting refugees were being given citizenship papers in exchange for political loyalty. It indicated demographic invasion and electoral manipulation without any proof to support the claim .

A YouTube video uploaded by a politically affiliated channel titled “One crore Rohingyas have entered India—wake up before it’s too late,” combined dramatic music with altered satellite imagery and archival footage. The video garnered over 12,000 views and was widely circulated via Facebook pages supporting nationalist causes. The caption warned of a “silent jihad,” likening the refugee influx to a planned insurgency.

While these narratives were less frequent than others, they generated high engagement and relied on the narrative of urgency and threat to culture. They were often framed as warnings or insider intelligence, using claims that mainstream media and authorities were covering up a demographic catastrophe to attract viewers.

b) Calls for harm reveal extreme hostility—and a shift from speech to incitement

Some of the most severe contents in the dataset went beyond just framing or stereotyping and explicitly called for violence against the Rohingya community. These “Call for Harm” narratives often took the form of brief but inflammatory commands, sometimes masked in nationalist rhetoric. In total, 30 posts were identified that explicitly called for harm— 29 on Facebook and 1 on YouTube.



In one Facebook post dated February 2025, a user wrote: “Country first—evict the Rohingyas now, or it’ll be too late.” Another stated, “Rohingyas must be kicked out before we can save Bangladesh.” Though phrased as civic calls to action, the posts used blunt, hostile language similar to the previous examples of incitement.

A particularly aggressive post read: “We will drag the Rohingya Muslims by their hair and throw them out.” It remained live for over 24 hours and by the time it was removed it had already

received dozens of shares. Another user wrote, “Time to cleanse Cox’s Bazar. Why are we feeding those who plan to destroy us?” Many of the “Call for Harm” posts were rated at the highest severity levels and frequently included references to armed retaliation or moral justifications for expulsion. Unlike the vague misinformation or fear driven narratives, they encouraged violence and even deemed it necessary.

These posts were typically shared by individual accounts, often with modest followings, but the intensity of their language and framing posed real risks for escalation, especially in an already volatile situation.

c) Posts falsely link Rohingyas to militant activity and portray the refugee population as a national security threat

On February 24, a video posted to Facebook claimed that “Rohingya militants” had attacked a Bangladeshi air force base. Within hours, the video was reposted by at least five other accounts and pages, accumulating over 75,000 views and over 2,000 reactions. The footage showed uniformed security personnel firing in an open area, but it showed no attackers, nor was the location identifiable as an air force installation. News reports later confirmed the incident involved a local dispute over a helmet at a checkpoint—unrelated to any Rohingya involvement ([The Business Standard, 2025](#)).

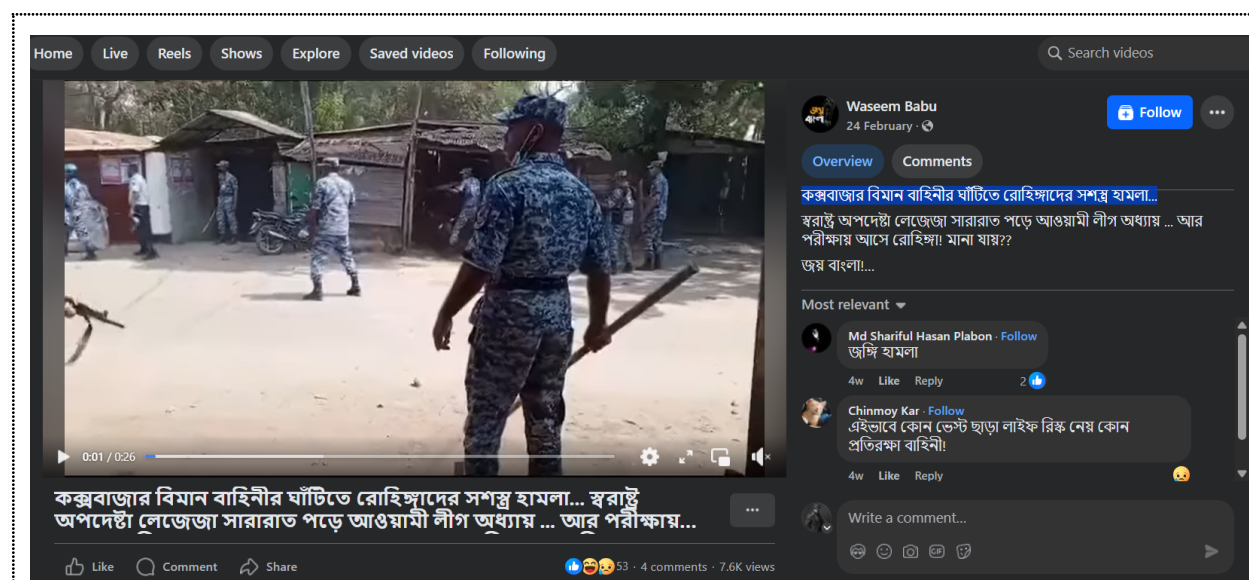


Image 03: False militant accusation spreads panic and criminalizes refugees

A screenshot of a Facebook post by an individual actor falsely claims, “Armed attack of Rohingyas against Bangladesh Air Force,” suggesting that the Rohingya community has engaged in militant aggression against state forces. This post, which received 53 likes and 4 comments, is a clear example of misinformation with geopolitical overtones, aiming to portray the Rohingya as a national security threat.

Despite this, the video and its reposts framed the event as a militant attack by the Rohingyas. Several posts used slogans aligned with the Awami League party and criticized the interim

government's security response. By portraying a local dispute as a Rohingya-led attack, the posts used misinformation to fuel political criticism, fears of militancy and state vulnerability.

Another similar post from December 2024, claimed that an armed Rohingya group is receiving jihadist training in Cox's Bazar. It showed masked individuals with firearms, later confirmed to be unrelated to the camps ([Fact-Watch, n.d.](#)). The word "terrorist" appeared 48 times across adversarial posts, reinforcing an ongoing narrative associating the refugee presence with national security threats.

d) Disinformation ties Rohingyas to Arakan Army, suggesting Bangladesh's territorial collapse

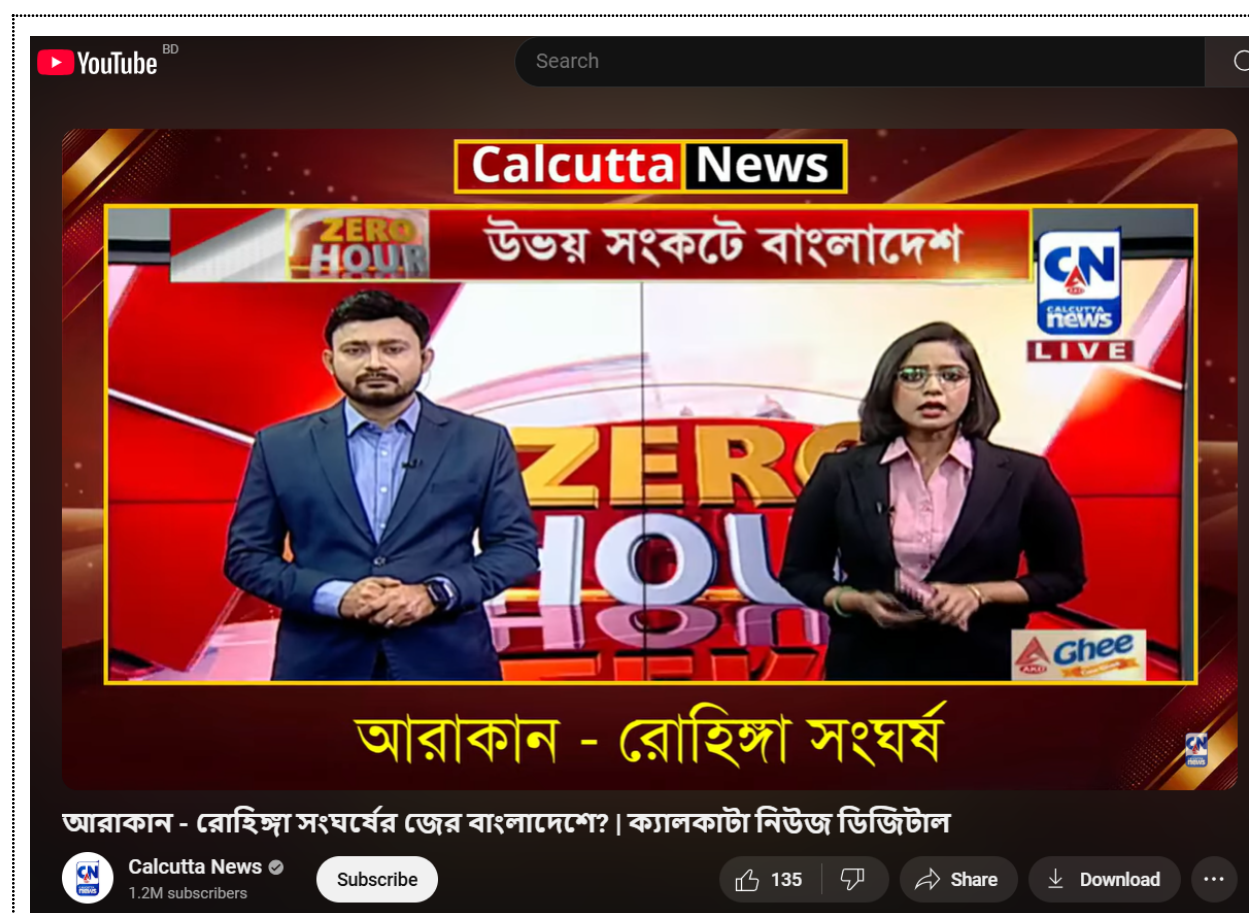


Image 04: YouTube narrative aligns Rohingyas with Myanmar junta and separatism

A YouTube video from a West Bengal-based media outlet claimed the Arakan Army was occupying parts of Bangladesh, while the Rohingya Salvation Army supported Myanmar's junta. With 10,000 views, the video framed the Rohingya as aligned with foreign military forces—fueling conspiracy-laden fears and portraying them as threats to national security.

During the monitoring period, a recurring disinformation theme claimed that the Arakan Army, an armed group based in Myanmar, was attacking or had already attacked Bangladesh. These

narratives linked regional conflict with the Rohingya presence and spread quickly across platforms.

A Facebook post from December 13 claimed, “There’s now a 100% doubt about Myanmar’s existence... The Arakan Army has taken over all bases... Repatriation of Rohingyas is no longer possible. America had asked Bangladesh to accept them as citizens—now we must listen.” Another post asserted that 262 kilometers of Bangladesh were under foreign control.

On YouTube, videos such as “Arakan Army invades Bangladesh” and “Chattogram may become independent by June 2025” relied on dramatic visuals, exaggerated claims, and conspiratorial hashtags. These narratives blended real conflict in Myanmar with false threats of secession and foreign manipulation, influencing viewers with high-production videos and political framing.

Posts in this category were frequently shared by financially motivated content creators and YouTube channels. Though fact-checked and debunked, they continued to circulate, showing how disinformation evolves through repetition and visual appeal.

e) Foreign conspiracy theories cast Rohingyas as agents of destabilization

A subset of adversarial content framed the Rohingya not as refugees but as means of foreign interference. These posts linked the refugee crisis to broader geopolitical rivalries involving the United States, China, and Pakistan.

Ateam · Follow
December 30, 2024 · 🌐

আমেরিকা শেখ হাসিনার কাছে কী চেয়েছিল?

বাংলাদেশে ২০১৮ সালের অনুষ্ঠিত সংসদ নির্বাচনের ফলাফল নিয়ে শুরু থেকেই দেশে-বিদেশে বড় ধরনের চাপে পড়ে শেখ হাসিনার আওয়ামী লীগ। একের পর এক গুজব, বিভ্রম, সমালোচনা আর নানা দাবীর মুখে গলদঘর্ম হতে হয় সরকারকে। দেশী-বিদেশী নানা ষড়যন্ত্র চলেতে থাকে জোরপোহে। জামাতে ইসলামী ও সমমনা উগ্রপন্থী ইসলামী দল, কওমি মাদ্রাসার শিক্ষক ও শিক্ষার্থী এবং বিভিন্ন নিষিদ্ধ ঘোষিত জঙ্গী সংগঠন আবারও সক্রিয় হয়। পাহাড়ের কুকি-চিন ন্যাশনাল ফ্রন্ট, আল-কায়েদা উপমহাদেশ শাখা, পুর... [See more](#)



882 47 comments 250 shares

Image 05: Screenshot of a viral Facebook post by 'Ateam' alleging U.S. support for Rohingya militants to pressure the Bangladesh government.

The conspiracy theory claims that the U.S. used the Rohingya crisis to pressure Sheikh Hasina's government, allegedly supporting armed Rohingya groups and backing separatist movements in southeastern Bangladesh. It suggests this was part of a broader plan to destabilize the region and force Bangladesh to align with U.S. geopolitical interests. This narrative was circulated through a Facebook page called "Ateam," which promotes the Awami League's perspective. The post gained traction with around 250 shares and over 800 reactions.

One post from December 31, 2024, alleged that the U.S. Burma Act was a cover for arming Rohingya-linked insurgent groups in southeastern Bangladesh. The post referenced military agreements, visa restrictions, and energy block negotiations, suggesting the Rohingya were pawns in a strategic campaign to pressure Bangladesh. Another post described the fall of the Hasina government as a Western-led "color revolution."

While speculative and unsubstantiated, these posts resonated with users by blaming foreign agenda for local challenges.

f) Suspicion toward NGOs is persistent, localized, and often tied to foreign interference narratives



Image 06: NATO aid withdrawal claim spreads anti-refugee misinformation

A screenshot of a Facebook post shared by an individual actor claims, “NATO has stopped all Rohingya funding for accusation of corruption,” a statement that is factually incorrect and misleading. The post, which received 214 likes, 11 comments, and 6 shares, is false. The narrative seeks to delegitimize aid efforts and reinforce negative perceptions of the Rohingya, fueling public skepticism and reducing sympathy toward their plight by falsely linking NATO to funding withdrawal over alleged corruption.

Throughout the period, posts repeatedly raised suspicion on NGOs operating in the camps. These posts rarely went viral but consistently created doubt by accusing aid agencies of mismanagement, profiteering, or sabotage.

One post from December 24, 2024, claimed that fires in the camps were timed to align with project deadlines—implying that agencies staged crises to secure funding. Another post alleged excessive presence of NGO staff, framing them as unaccountable outsiders. A third criticized the employment of foreign nationals in senior roles at international agencies, linking aid operations to loss of work for locals.

NGOs were also mentioned in visual disinformation, including a repurposed image falsely claiming that NATO had withdrawn all humanitarian aid. Although the image showed a 2017 visa waiver signing between Rwanda and the UAE, the new caption turned it into a declaration of abandonment by global institutions.

g) Disinformation frame Rohingyas, particularly women, as a public health risk

Another bunch of posts portrayed Rohingya refugees as a threat to public health and social morality. Several Facebook posts linked the spread of HIV/AIDS in Bangladesh to Rohingya women, often using vague or inflated figures without any credible sources.

A widely shared post warned that more than 900 Rohingya women were HIV-positive and involved in sex work. The post blamed both foreign-returned Bangladeshis and refugees, reinforcing stereotypes of moral decay and external influence.



Image 07: Mainstream media post falsely links Rohingyas to AIDS crisis

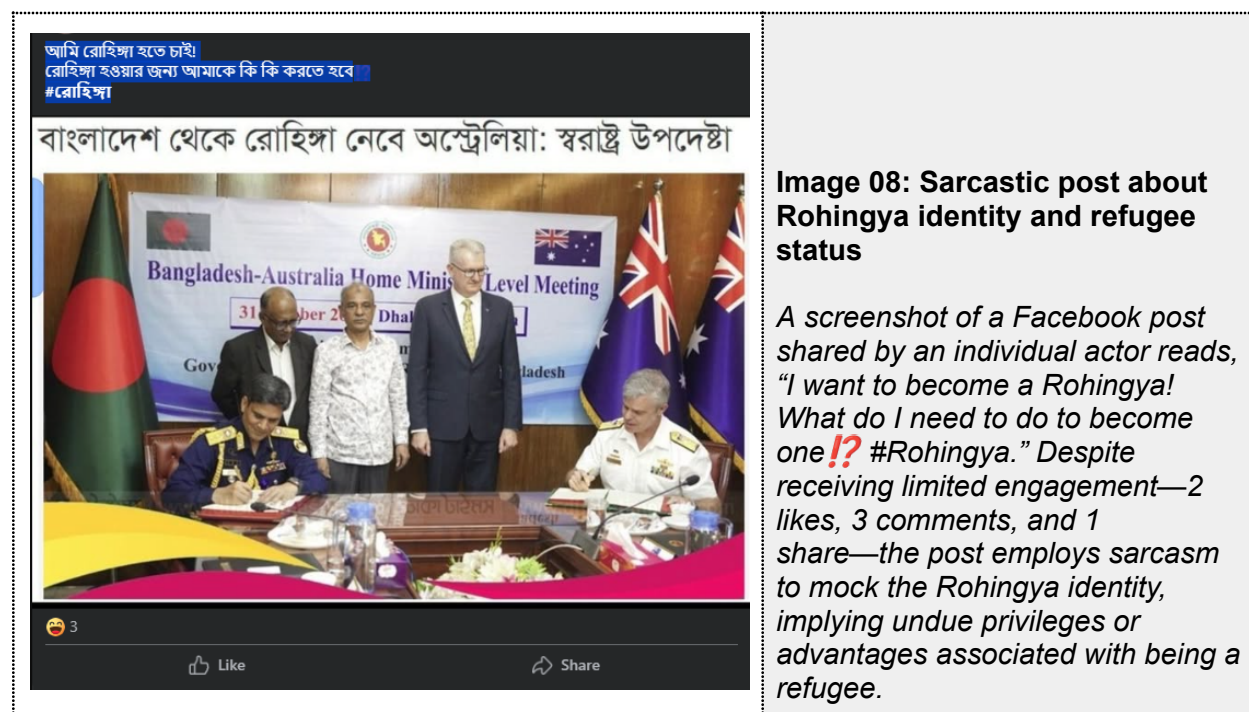
A widely circulated Facebook post by a Bangladeshi media outlet claimed, “Due to the Rohingyas, AIDS cases are rising alarmingly,” using hashtags like #Rohingyas and #AIDS to boost reach. With 51,000 views and nearly 3,000 likes, the post spread a sensational, stigmatizing narrative that falsely linked refugees to a public health crisis.

Two national news outlets amplified these claims. RTV released a Facebook video headlined "Due to Rohingyas, AIDS patients are increasing alarmingly," which received over 2,900 reactions and Independent24.tv followed with a similar report. Both pieces presented health data without proper sources, thus further spreading stigma and misinformation.

Another set of posts focused on reproductive behavior, alleging that Rohingya families had up to 17 children signifying a strain on public resources. These narratives exaggerated fertility rates to question the community's ability to integrate, often casting reproduction as a deliberate burden on the host country.

h) Sarcasm turns exclusion into everyday expression

Some of the most pervasive narratives in the dataset were not overtly hateful but relied on subtle sarcasm and comparative framing to create distance between the host population and the Rohingya. These posts reinforced the idea that the Rohingya are outsiders receiving undue privileges or causing local neglect.



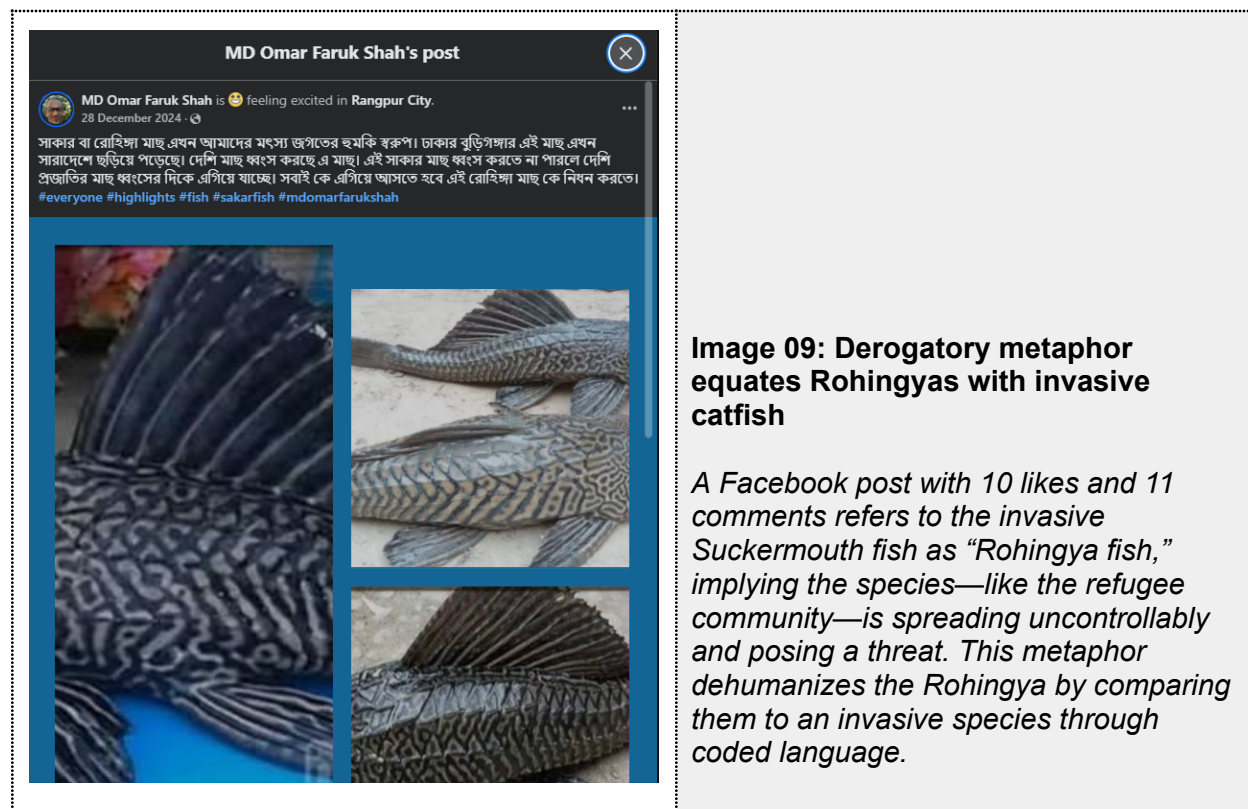
Several users used sarcasm to suggest that life as a Rohingya in Bangladesh was easier or more privileged than for local citizens. In one widely shared post, a user joked that given the current conditions, it might be better to stay in the country “as a Rohingya” rather than attempt to migrate. Others expressed frustration by comparing the term “Rohingya” to a curse word, implying that it now symbolized neglect or inferiority in public discourse.

Such comparisons reflected broader resentment that government services, humanitarian aid, or even bureaucratic procedures were more accessible to refugees than to citizens—whether or not those claims were grounded in reality. The term “Rohingya” had become a tool for expressing dissatisfaction with governance, red tape, and inequality—at the expense of an already vulnerable population.

i) “Rohingya” as a metaphor turns identity into insult

Beyond sarcasm or policy critique, many posts in the dataset used the term “Rohingya” as a metaphor for illegality, disorder, or inferiority. These posts don’t just target the community—they turn the identity into a slur.

In some examples, the term “Rohingya” was used to describe disorganized political movements or internal party conflicts, implying infiltration or collapse. Other posts extended this metaphor to non-human subjects: auto-rickshaws without proper licenses were labeled as “Rohingya CNGs,” and invasive catfish species in the Buriganga River were mockingly dubbed “Rohingya fish.”



These comparisons reframe the community name into a symbol of illegitimacy, lawlessness, and threat. By turning the word “Rohingya” into an insult or a punchline, such narratives reinforce the idea that their presence—like that of pests or unregulated vehicles—is unnatural and undesirable.

It reveals how language itself becomes a tool for harm and the weaponization of language, layered into jokes, comparisons, and memes, contributes to the everyday normalization of exclusion.

j) Positive stories may be fewer—but they carry the potential to connect, affirm, and shift the narrative.

Amid a flood of harmful content, the study captured 233 posts that took a different path—sharing stories of support, resilience, and shared humanity. While smaller in number, these posts offered a powerful countercurrent to hostility, reminding us that not all public sentiment is driven by fear

or exclusion. On facebook, such posts garnered 272 reactions on an average, more than several adversarial narrative categories.

Some of the most consistent messages of solidarity came from institutional sources. Posts highlighted UN agencies and NGOs—like UNHCR, UN Women, and Oxfam—for their work in providing vocational training, community services, and skill-building opportunities. These messages weren't always dramatic, but they reflected a steady engagement from viewers.

Moments of shared joy and connection also surfaced—such as football or kabaddi tournaments between Rohingya and host community youth. These snapshots offered rare glimpses of mutual visibility, reminding that coexistence isn't just possible—it's already happening in small but meaningful ways.

After a major fire destroyed 600 homes in the camps, many online users responded with concern and empathy, reflecting a human instinct to care, even in polarized environments. Similarly, stories about international diplomatic visits or passport policy changes triggered constructive dialogue about refugee rights and state responsibilities.

Unexpected voices joined this space too. A large gathering of Islamic scholars and Olama groups issued public support for the Rohingya cause, blending religious language with humanitarian values—an unusual but powerful moment that showed how advocacy can emerge from many corners.

7. How actors shape narratives

This chapter examines the key actors responsible for generating and amplifying narratives about the Rohingya across social media platforms. Drawing from data collected between December 2024 and February 2025, it identifies three main actor types—media organizations, political entities, and individuals—and analyzes their volume of activity, thematic focus, and engagement levels.

Comparative insights: Who drives what?

- **Misinformation:** Individuals contributed to 85% of misinformation-tagged posts, often involving false claims about documents, benefits, or foreign conspiracies. Political actors contributed just 2%, but those posts had the highest per-post engagement.
- **Disinformation:** Media actors accounted for 21% of all disinformation posts—mostly through exaggerated or misleading framing, especially in Indian outlets. Individuals still take the lead in total volume.
- **Hate speech:** Political actors had the highest proportion of hate speech within their limited output (4 out of 8 posts). Media actors were also significant here, with 32% of their adversarial content qualifying as hate speech, especially via slurs and fearmongering. Individuals made up the rest, frequently using stereotyping or exclusionary language.

7.1 Media actors: Fewer posts, outsized reach

Media outlets—though responsible for just 6.2% of all content in the dataset—played a disproportionately influential role in shaping public sentiment about the Rohingya. On Facebook, media pages received an average of 879 reactions per post, while on YouTube their content averaged over 7,125 views per video—the highest among all actor types. This suggests that media-driven narratives travel further and garner more attention, despite being a minority in volume.

a) Indian media channels pushing anti-Rohingya narratives across platforms

Indian outlets such as Republic Bangla, ABP Ananda, and Aaj Tak Bangla emerged among the most active and influential. They frequently published content that framed the Rohingya as a security threat, linking them to drug trafficking, terrorism, or demographic invasion. These narratives typically relied on negative targeting and in-group/out-group framing, sometimes wrapped in sensational storytelling.

Examples:

- A YouTube video from Republic Bangla claimed: “Rohingyas are now roaming in India with weapons. Who will take responsibility?”

- ABP Ananda pushed similar claims on Facebook, suggesting a Rohingya connection to organized crime, without citing verifiable sources.

These narratives feed into cross-border anti-Muslim and nationalist discourses already present in Indian media ecosystems. Despite the international origin of the content, this material often circulates widely in Bangladesh’s online spaces, amplifying xenophobia.

Outlets like BBC Bangla and Al Jazeera Bangla appeared less frequently in the dataset but tended to offer more balanced or neutral reporting. Still, some of their content was shared with hostile framing by third-party actors, distorting the original message.

International media represented 44% of all media-tagged content and contributed a large share of total views on YouTube. Their average post had 7,637 views—more than any other category.

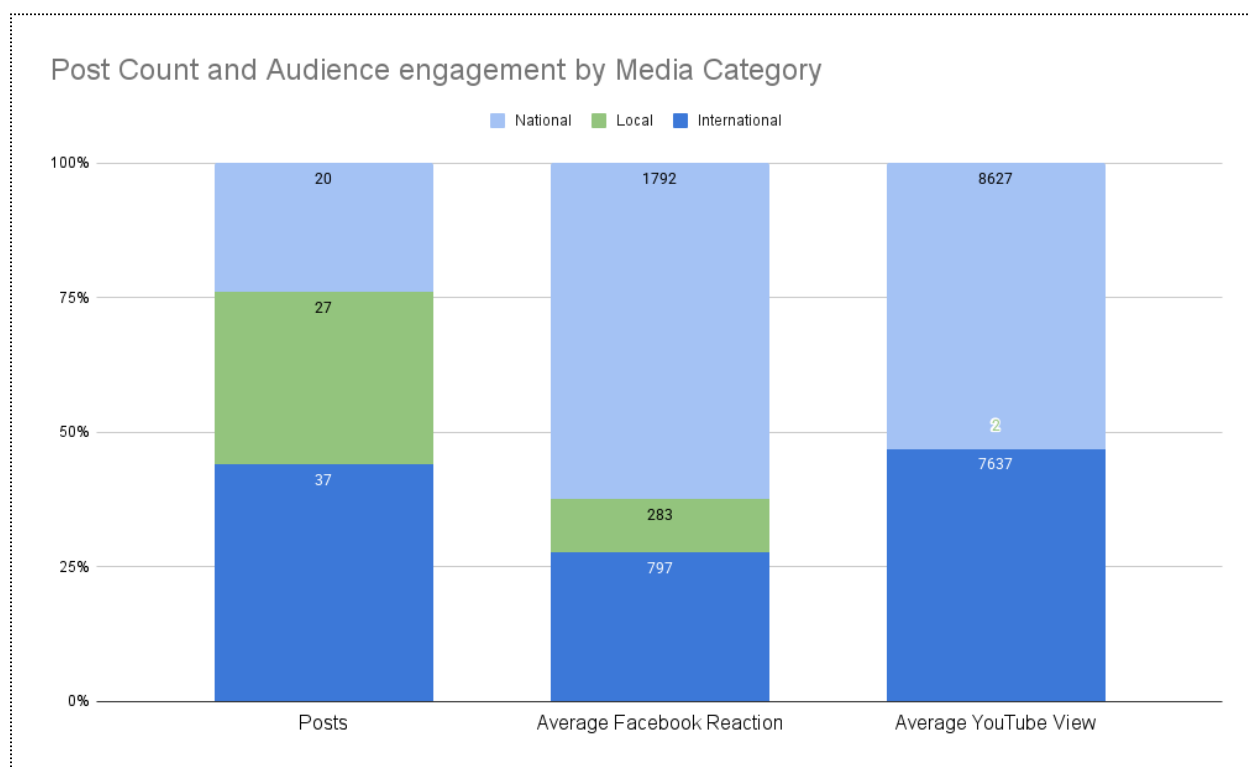


Chart 11: Comparison of post count, average Facebook reactions, and average YouTube views across international, local, and national media platforms.

b) Local media outlets from Cox’s Bazar driving high engagement

Local news pages such as Channel Ukhiya, UkhiyaNews.com, and Teknaf 71 may have posted fewer videos or articles than their Indian counterparts, but their posts drove high engagement—especially on Facebook. On average, posts from national Bangladeshi outlets received 1,792 reactions per post on Facebook, while local outlets received 283.

Their narratives were often emotional, sensational, or suggestive—focusing on rising crime, deteriorating law and order, or “ungrateful” refugee behavior.

For Example:

- One Facebook post from Channel Ukhiya featured an image of a group of men near the camps with a caption: “The government feeds them, and now they rob our houses. Who is responsible?”
- Teknaf 71 ran stories with titles like “Rohingyas now armed in the hills” or “Rohingya youth caught with yaba”—relying heavily on criminalization narratives.

c) National media: Fewer posts, louder impact

National outlets, such as Ekattor TV, The Daily Star, DBC News, and Bangladesh Pratidin, were responsible for 23.8% of all media-tagged posts. Yet their posts generated the highest average Facebook engagement: 1,792 reactions per post—well above international (797) and local (283) media.

This suggests that when national media publish content related to the Rohingya, it resonates more widely across general audiences in Bangladesh, potentially influencing middle-class, urban sentiment.

While some national outlets maintained a more professional tone, others echoed security concerns and mirrored government positions. Their content frequently emphasized negative targeting and criminalization. Half of national media’s adversarial posts used negative targeting. Local media were more prone to sensationalism, which appeared in 37% of their adversarial content.

For Example:

- A Facebook post by Ekattor TV used the headline: “Rohingyas not only fleeing camps but forming gangs in Chattogram hills.”
- DBC News ran a story alleging Rohingya links to a local murder case, using speculative language and unnamed sources.

In essence, national media serve as a middle force—less extreme in tone than many local or Indian pages, but still capable of shaping dominant narratives, especially when covering crime, security, or camp unrest. Their relatively infrequent coverage may give the impression of neutrality, but when they do publish, the framing often reinforces fears, subtly legitimizing exclusionary narratives.

Top 5 Media actors by adversarial post		
Name	Post count	Category
Republic Bangla	18	International
Aaj Tak Bangla	6	International
Channel Ukhiya	3	Local
ABP Ananda	3	International
UkhiyaNews.Com	2	Local

7.2 Political actors: Small footprint, outsized tone

Political organizations accounted for just 0.6% of all monitored posts—only nine out of 1,297 across Facebook and YouTube. Yet despite this limited presence, their content stood out for its intensity and the nature of narratives employed.

The shift in the online landscape following the political transition in Bangladesh on August 5 is also significant. During the baseline period, hundreds of pages—some clearly linked to political parties or state entities—actively spread misinformation, disinformation, or hate speech about the Rohingya. After August 5, many of these pages went offline or ceased activity, reducing overt political participation in MDH narratives.

All political actor posts were found on Facebook, and none on YouTube during the monitoring period. Of these, the dominant narrative types included negative targeting and in-group/out-group framing—messaging that depicted the Rohingya community as outsiders or threats to national identity. These posts often leaned on adversarial tropes like ridicule, exclusion, or coded calls for separation.

Though small in number, the engagement figures tell another story. Posts from political organizations received the highest average reactions per post on Facebook—1,428, far exceeding individual (87) or media actors (865). One post tagged as misinformation, in particular, drew over 9,000 reactions. This suggests that while political actors were less active, the few messages they did share traveled widely.

Though political actors accounted for less than 1% of the total content, their posts were often sharply worded and intended to frame the Rohingya in hostile and conspiratorial terms. These posts mostly originated from Bangladesh and India, reflecting local and cross-border political interests.

For Example:

- Bangladesh Awami League (Facebook): Posted content framing Rohingya refugees as a threat to national stability. One post, labeled as misinformation, discussed the Rohingya camp fire with the caption: “Rohingya camp fire: look at the calculation! Find out who benefits.”
- Bangladesh Students’ League (Facebook): Focused on the geopolitical implications of the Rohingya presence, suggesting that foreign powers like the US and China are manipulating the refugee issue.
- Aam Aadmi Party Uttar Pradesh (Facebook, India): Shared disinformation suggesting Bangladeshi Rohingyas are infiltrating India. One example read: “How did the Rohingya of Bangladesh come to India and get Aadhaar?”
- BJP Tripura (Facebook, India): Accused opposing parties of supporting Rohingyas for political gain: “AAP-Da (Apat) love with Rohingyas and Bangladeshis is not hidden.”
- Romjan Nagar Union Krishok Dol and Keraniganj Model BNP: Posted content that framed Rohingyas as criminals or infiltrators. One post read: “This Rohingya Jalil has been brokering Awami League leaders for years.”

Political narratives reflect a pattern of in-group/out-group framing and negative targeting. Despite their low volume, posts from political actors had significantly higher average reactions on Facebook—1,428 per post, compared to 87 for individuals and 865 for media. This suggests that their content, while sparse, can have considerable impact.

Top 5 political actors by adversarial post		
Name	Facebook Reaction	Country
Bangladesh Awami League	9191	Bangladesh
Aam Aadmi Party Uttar Pradesh	2512	India
Bangladesh Students' League	874	Bangladesh
BJP Tripura	106	India
Trinomul Awami League Jhenaidah	97	Bangladesh

7.3 Individual actors: High volume, complex dynamics

Individual users produced the overwhelming majority of posts in the dataset—accounting for more than 90% of all content. This group, however, was highly diverse. It included everyday users, religious commentators, financially motivated content creators, and individuals known for prior disinformation campaigns.

Some users appeared affiliated with political parties or local leadership circles in Bangladesh and India. However, given the volume and often opaque nature of profiles, it was methodologically impractical to reliably establish political alignment for each user. Due to the complexity and limitations of the study scope, we opted to retain only those political affiliations explicitly mentioned, rather than inferring alignment from indirect cues.

Narratives included exaggerations of Rohingya criminality, demographic fears, conspiracies about foreign support, and calls for repatriation or exclusion. Unlike media or political actors, individuals often employed cruder language and emotional appeals. Yet their scale made them critical to the amplification of adversarial discourse.

For Example:

- One user wrote, “Rohingyas are like carbon monoxide—dangerous but unnoticed.”
- Another said, “Cut hills for Rohingyas, demolish houses of locals. This is Bangladesh.”
- A viral meme declared, “A day will come when Rohingyas will rule over Cox’s Bazar.”

The data shows that individuals were especially active in sharing stereotyping (42% of their adversarial posts) and fearmongering narratives (30%), often laced with moral outrage or sarcasm. Their role underscores how adversarial narratives do not always require coordinated campaigns—they thrive when enough individuals repeat them consistently.

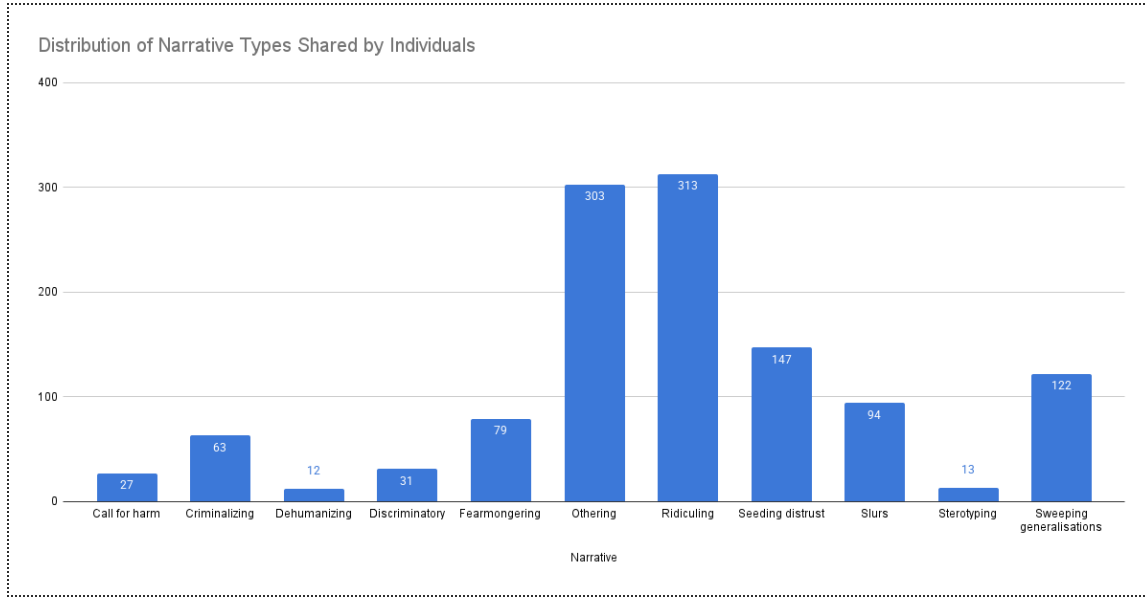


Chart 12: This chart illustrates the frequency of various narrative types shared by individual entities. "Ridiculing" and "Othering" narratives are the most prevalent, followed by "Seeding distrust" and "Sweeping generalisations".

While most individuals posted a handful of times, a few stood out either by volume or reach:

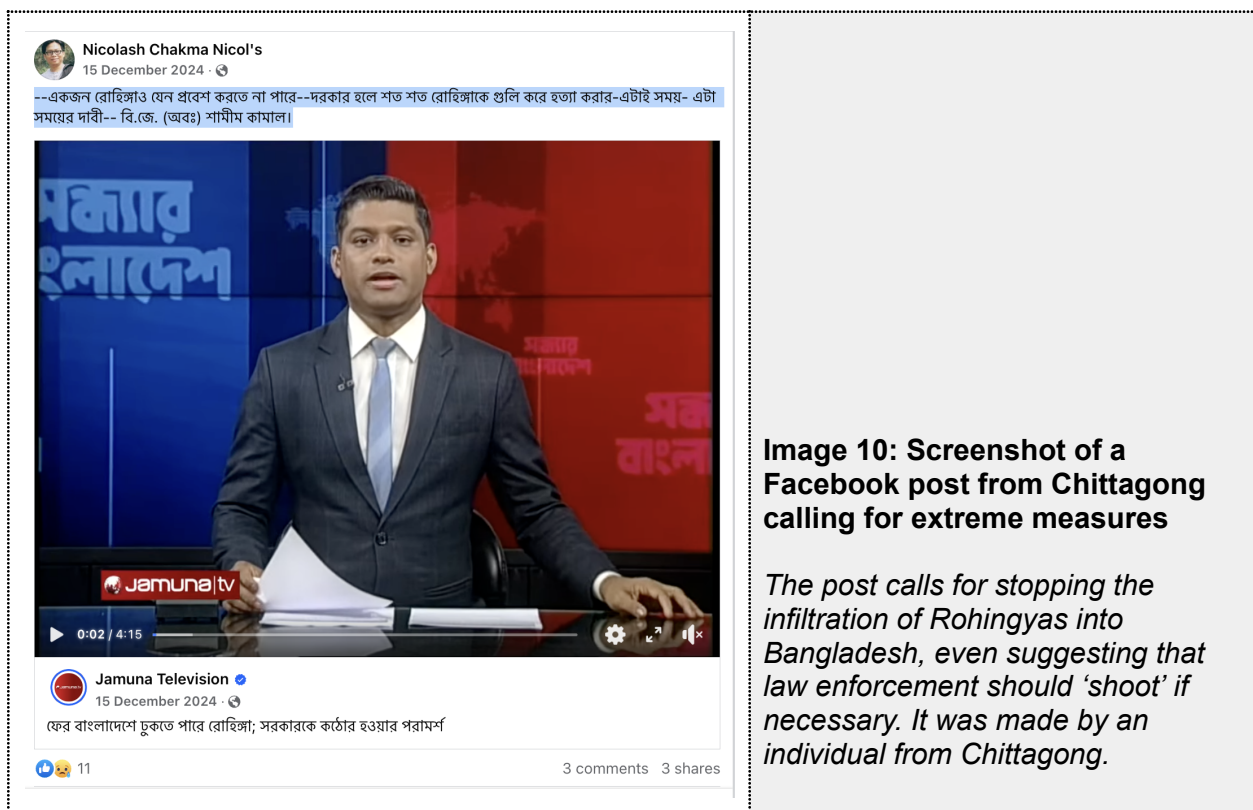
- Mohiuddin Jewel (Facebook) was the most active individual user, posting 18 times during the monitoring period. His posts often repeated themes of Rohingya criminality and the need for repatriation.
- Mohammad Faisal (Facebook) posted 13 times with several posts containing misinformation, including one falsely claiming Rohingyas had received Bangladeshi passports.
- NewsBDTime (YouTube) had the highest average views per video (20,511), despite posting fewer than 10 times. Their content was generally framed around national security threats.
- Noyon Baul (YouTube) received the most total views (94,000), mostly on one viral video suggesting a Rohingya conspiracy to overrun Chattogram.

Top 5 Individual actors by adversarial post	
Name	Post count
Mohammad Saiful Islam	9
Norul Bashar (নূরুল বশর)	5
Rohingya: Silent Victims of Injustice	4
JM Salim	4
Somen Pandey	3

8. Platform accountability and policy enforcement

This chapter examines how well Facebook and YouTube enforce their stated policies on hate speech, misinformation, and harmful content in the context of narratives targeting the Rohingya community. Although both platforms maintain detailed policies and claim to use AI-driven tools, human moderation, and fact-checking partnerships, the dataset reveals significant enforcement failures in Bangla-language content.

Despite clear policy guidelines, harmful content often goes unmoderated



Both Facebook and YouTube maintain explicit policies prohibiting hate speech, particularly content that dehumanizes, incites violence, or discriminates based on identity. Yet, data from the monitoring period revealed dozens of posts that clearly violated these guidelines but remained online without visible moderation. Many were widely shared, attracting high engagement across both platforms.

Among the clearest examples were direct calls for violence:

- “Not a single Rohingya should be allowed in—if necessary, hundreds of them should be shot dead.”

Such content qualifies as hate speech under Meta’s prohibition on “calls for violence or serious harm” and YouTube’s ban on inciting hatred based on identity. Nonetheless, many of these posts remained visible for hours—or in some cases, days—suggesting lapses in enforcement.

Slurs and metaphors dehumanize Rohingyas— escape platform response

Hate speech also appeared in more coded or metaphorical forms, especially through comparisons that stripped the Rohingya of human dignity. A number of posts used slurs or analogies to describe the community:

- “These frauds—these Rohingyas who claim citizenship rights in Bangladesh—are dogs.”
- “The Rohingyas are like carbon monoxide—dangerous but unnoticed.”



Image 11: Screenshot of a YouTube video spreading false information

The video falsely claims that one crore Rohingyas have entered West Bengal, India—a claim not supported by any credible data. It was spread through a YouTube channel with over one million subscribers.

These examples fall within Facebook’s rules against dehumanizing language and YouTube’s restrictions on slurs that incite hatred or demean protected groups. Yet, several of these posts

remained live and were widely circulated. In one case, the phrase “drag the Rohingyas out by their hair” was republished in multiple forms across West Bengal-based channels.

Misinformation spreads unchecked on YouTube, absent third-party verification

While Facebook partners with fact-checking organizations in Bangladesh, YouTube has no such local partnerships in place. This gap was evident in the persistence of misinformation narratives that violated both platforms’ stated policies. One widely shared video falsely claimed:

- *“One crore Rohingyas have entered India—wake up before it’s too late.”*
- *“Nearly 10 million Rohingya Muslims have entered West Bengal, altering the demographic balance. The Muslim population now exceeds 40%.”*

Do these posts cause egregious harm that YouTube’s guidelines demand for taking any action? It needs nuanced assessment. But such claims—lacking any factual basis—stoked fear and xenophobia, and continued to receive thousands of views.

Fact-checkers in Bangladesh had previously debunked several of these narratives, including claims linking Rohingyas to national security threats or foreign conspiracies. However, in the case of Facebook, most of these claims were not flagged at the time the monitoring team accessed them. According to Meta’s policy, content rated as false by its third-party fact-checking partners should be labeled and have its distribution reduced. The continued visibility of these posts suggests limitations in the enforcement of this policy or delays in its application.

Moderation gaps widen in low-resource languages and across borders

Many of the most harmful posts in the dataset were written in Bangla, including posts targeting audiences in West Bengal, India. These posts used similar narratives to those found in Bangladesh—focusing on demographic panic, national security, or political conspiracy—but were often framed with regional references or targeted at Indian elections.

For example:

- *“West Bengal will be erased from India’s map... If this continues, West Bengal will soon become Bangladesh.”*
- *“The time has come—drag the Rohingyas out by their hair! Hindus must be protected in Bengal.”*

Such posts spread across multiple Facebook pages and YouTube channels, often linked to political groups or regional influencers. The cross-border and multi-platform nature of these narratives made enforcement more complex. Facebook and YouTube policies cover the same protected characteristics, but enforcement appeared inconsistent across geographies and actor types.



Moreover, the reliance on automated moderation and community reporting likely underperformed in Bangla, a low-resource language for which detection tools remain underdeveloped. This allowed hate speech and disinformation to circulate widely before drawing attention.

Repeated violations, limited consequences

Finally, a small number of accounts were responsible for repeatedly sharing content that clearly violated platform rules. Some accounts shared identical hate speech posts across multiple pages, while others reposted debunked misinformation narratives weeks after they had been fact-checked. Despite this, these accounts continued to operate openly.

In the dataset, multiple Facebook pages reposted the false claim that Rohingya militants had attacked a Bangladeshi air force base—despite the incident being verified as a local clash unrelated to the refugee population. It indicates gaps in platform responses to recurrent bad actors, and automated detection of already flagged misinformation content.

9. Strategic recommendations

Online hostility against the Rohingya is often the most persistent—and the most visible. Who's behind this? In most cases, it's everyday people. Individuals made up over 90% of all posts—sharing exaggerated fears, stereotypes, and sometimes outright calls for harm. But when media or political actors entered the fray, their messages traveled faster and landed harder. It makes the online space more unsafe for the refugee community.

Based on what we've learned, the following recommendations offer possible ways forward—not as blueprints, but as building blocks for deeper engagement, grounded in collaboration with Rohingya communities and host society actors alike.

1. **Invest in refugee-sensitive journalism—and the editors behind it**

Media pages made up just 6.2% of all content but generated the highest per-post reach. That level of visibility comes with responsibility. Often, the best way to tackle misinformation or disinformation is with facts and sensitive portrayals of the affected and the vulnerable. UNHCR and civil society partners should engage local and national newsrooms with training, editorial exchanges, and guidelines on reporting in conflict and refugee contexts. News leaders, not just reporters, should be part of these efforts. Ethical frameworks, real-world case studies, and regional peer learning are crucial to shifting the tone and framing in high-engagement news environments.

2. **Build media literacy that goes beyond classrooms**

The data shows a clear pattern: individual users are the biggest drivers of harmful content, especially stereotyping and fearmongering. But they're also the least equipped to question it. Media literacy programs must move outside formal education and into communities—mosques, schools, village committees, Facebook groups. Add sensitivity toward conflict-affected communities to the curriculum. Make it widely available by engaging political and social actors. UNHCR can support localized Bangla- and Rohingya-language materials and work with platforms to fund co-branded initiatives that build critical thinking at scale. Media literacy that is refugee-sensitive may require contextual understanding and expertise. It cannot be generic.

3. **Set up a trusted channel with platforms for faster response**

Too many posts flagged as hate speech or misinformation stayed up—especially in Bangla. Many had already been fact-checked and debunked. UNHCR and civil society organizations should have a direct line of communication with content moderation teams at Meta, YouTube, and others. Build escalation channels independently and through the Trusted Partner networks of social media platforms. These partners can serve as trusted flaggers, helping platforms act quickly on dangerous content—especially when automated detection fails. Often, it will require looking beyond a piece of content to assess the systemic risks it poses to disenfranchised people.

4. **Track shifts in harmful narratives with better monitoring tools**

After the political shift in August 2024, official party-linked pages largely went silent—but individuals filled the gap. Hate didn't disappear; it just changed hands. To keep up, there needs to be consistent, multi-language monitoring of trends, narratives, and actors. A

cross-platform early warning system—tracking Bangla, Banglish, and Rohingya—can help UNHCR and its partners respond faster and more accurately, especially during moments of political volatility or crisis.

5. Push for greater transparency in algorithms and moderation

High-severity hate speech—especially on YouTube—received the highest average views of any content type, despite violating community standards. UNHCR should work with international advocacy networks to demand better algorithmic transparency and targeted investment in Bangla-language moderation, including calls for more human moderators, clearer appeal processes, and stronger penalties for repeat violators.

6. Support cross-border coalitions and Rohingya-led research

Some of the most damaging narratives didn't start in Bangladesh—they came from India, amplified by political actors and fringe media pages with regional agendas. That's why responses must go beyond borders. UNHCR should fund cross-country research collaborations, support Rohingya-led digital rights work, and back advocacy efforts that address shared harms across language and geography. Evidence is key—and at scale. It may help inform policies and change enforcement strategies.

7. Back content creators who challenge harmful narratives

Some of the most powerful counters to hate are creative, community-rooted, and persuasive—not just factual. But positive content often lacks visibility. Don't ignore the power of positive messaging. The study found 233 positive posts about the Rohingya—more than the number of hate speech posts. UNHCR and partners can support local journalists, YouTubers, and artists who want to share alternative stories—about coexistence, justice, or shared struggle. This means funding, training, and network-building to help positive content compete in the same viral spaces where disinformation thrives.

8. Create space for public conversation—not just monitoring

Monitoring MDH content is essential, but it's not enough. We need spaces where affected communities—Rohingya refugees, host community members, young people, religious leaders—can talk, debate, and reflect on what they see online. UNHCR should convene regular community dialogues, online or offline, where people can unpack viral narratives and explore shared concerns. These spaces can help reduce polarization and build empathy—long before policies or enforcement catch up.

These recommendations are not a checklist. They are a reflection of patterns that need attention—but any action must be co-created with those affected most. Countering hate and disinformation against the Rohingya cannot succeed without the full participation of refugee communities, local actors, platforms, journalists, civil society and political actors among others. A human rights-based and stakeholder-driven approach is the sustainable path forward.

References

- UNHCR. (n.d.). Rohingya refugee crisis explained. <https://www.unrefugees.org/news/rohingya-refugee-crisis-explained/>
- International Organization for Migration. (n.d.). Rohingya response. <https://www.iom.int/rohingya-response>
- Council on Foreign Relations. (2020). The Rohingya crisis. <https://www.cfr.org/backgrounder/rohingya-crisis>
- UNHCR. (n.d.). Rohingya emergency. <https://www.unhcr.org/us/emergencies/rohingya-emergency>
- Human Rights Watch. (2021, June 7). An island jail in the middle of the sea: Bangladesh's relocation of Rohingya refugees to Bhasan Char. <https://www.hrw.org/report/2021/06/07/island-jail-middle-sea/bangladeshs-relocation-rohingya-refugees-bhasan-char>
- Ansar, A., & Khaled, A. F. Md. (2021). From solidarity to resistance: host communities' evolving response to the Rohingya refugees in Bangladesh. *Journal of International Humanitarian Action*, 6(1). <https://doi.org/10.1186/s41018-021-00104-9>
- France. (2022, October 20). Welcome no more: Rohingya face backlash in Bangladesh. France 24. <https://www.france24.com/en/live-news/20221020-welcome-no-more-rohingya-face-backlash-in-bangladesh>
- The Daily Star. (2023, August 21). Refugee camps: Killings, narco crimes on the rise. <https://www.thedailystar.net/news/bangladesh/crime-justice/news/refugee-camps-killings-narco-crimes-rise-3402601>
- Khatun, A. (2024). Media, propaganda, and the othering process of the Rohingyas. In *International perspectives on migration* (pp. 169–199). https://doi.org/10.1007/978-981-97-1424-7_11
- Reuters. (2018, March 6). Facebook's role in Rohingya crisis under scrutiny. <https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUKKCN1GO2PN>
- Tähtinen, T. (2024). When Facebook is the Internet: The role of social media in ethnic conflict. *World Development*, 180, 106633. <https://doi.org/10.1016/j.worlddev.2024.106633>
- Schissler, M. (2024). Beyond hate speech and misinformation: Facebook and the Rohingya genocide in Myanmar. *Journal of Genocide Research*, 1–26. <https://doi.org/10.1080/14623528.2024.2375122>
- Bangladesh Telecommunication Regulatory Commission. (n.d.). About BTRC.

<https://btrc.gov.bd/site/page/347df7fe-409f-451e-a415-65b109a207f5/->

Kemp, S. (2025, March 3). Digital 2025: Bangladesh — DataReportal – Global Digital Insights. DataReportal – Global Digital Insights.

<https://datareportal.com/reports/digital-2025-bangladesh>

Meta. (n.d.). Community standards. <https://transparency.meta.com/policies/community-standards/>

YouTube. (n.d.). Community guidelines.

<https://www.youtube.com/howyoutubeworks/policies/community-guidelines/>

Heath, E. C. a. B. (2021, October 26). Language gaps in Facebook's content moderation system allowed abusive posts on platform: Report - The Wire. The Wire.

<https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts>

Faria, F. T. J., Baniata, L. H., & Kang, S. (2024). Investigating the Predominance of Large Language Models in Low-Resource Bangla Language over Transformer Models for Hate Speech Detection: A Comparative Analysis. *Mathematics*, 12(23), 3687. <https://doi.org/10.3390/math12233687>

Amnesty International. (2023, October 31). Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report.

<https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/>

Chimkhilai, P., & Panyametheekul, S. (2024, January 31). Hate speech in YouTube comments on Rohingya refugees in Thailand and Syrian refugees in Europe.

<https://so04.tci-thaijo.org/index.php/LEARN/article/view/270379>

The Daily Star. (2023, December 8). Meta removes 50 FB accounts, 98 pages linked to AL, CRI. The Daily Star.

<https://www.thedailystar.net/news/bangladesh/news/meta-removes-50-fb-accounts-98-pages-linked-al-cri-3623276>

Google Transparency Report. (n.d.).

<https://transparencyreport.google.com/youtube-policy/removals?hl=en>

Misinformation on YouTube: High profits, low moderation. (2024, October 1). Dismislab.

<https://en.dismislab.com/misinformation-on-youtube-high-profits-low-moderation/>

Report, T. (2025, February 25). ISPR says Air Force base attacked in Cox's Bazar; 1 shot dead during clashes. The Business Standard.

<https://www.tbsnews.net/bangladesh/ispr-says-air-force-base-coxs-bazar-attacked-police-confirmed-1-dead-1076756>

মিয়ানমারের সশস্ত্র মুসলিম গোষ্ঠীর ভিডিওর সঙ্গে বাংলাদেশকে জড়িয়ে বিভ্রান্তিকর প্রচারণা | Fact-Watch. (n.d.).

Fact-Watch. <https://www.fact-watch.org/myanmar-armed-group-misleading-bangladesh/>

Annex 1

Adversarial narrative Framework

In this study, we categorize online narratives concerning the Rohingya community using a framework adapted from the Global Disinformation Index's adversarial narrative model. We distinguish adversarial narratives into three main types: Negative Targeting (which includes demeaning language such as slurs, ridicule, and content that seeds distrust), Sensational Content (which amplifies emotional responses through sweeping generalizations and coded language), and In-Group/Out-Group Framing (which positions the Rohingya as perpetual outsiders through dehumanization, criminalization, and calls for harm). In addition, we differentiate content into two primary types—misinformation and hate speech—and identify various actors (individuals, media, political, state, and religious organizations) who contribute to these narratives.

Adversarial Narratives: According to GDI, adversarial narratives are "rooted in, involve, or are strongly characterised by conflict or opposition between actors and their interests, and especially between a social in-group and an out-group. When adversarial narratives are deployed, they create a series of smaller conflicts played out across the internet." These narratives are effective because they "inflame social tensions by exploiting and amplifying perceived grievances of individuals, groups, and institutions." We categorized the Rohingya related adversarial narratives in three types:

In our analysis, we categorized adversarial narratives into three primary types, each further divided into subcategories:

1. Negative Targeting: This category includes narratives that use slurs, microaggressions, or mocking language to ridicule, create distrust, and incite fear against the Rohingya community.

- Sarcastic/Ridiculing: Posts mocking or belittling the Rohingya community, often employing humor or exaggerated statements to demean them.
- Slurs/Microaggressions: Subtle or explicit offensive terms and remarks that reinforce prejudice against the Rohingya.
- Seeding Doubt/Distrust: Content designed to create suspicion about the Rohingya, portraying them as untrustworthy or harmful to societal stability.
- 2. Sensational Content: Sensational narratives amplify hostility by exaggerating claims or triggering emotional responses against the Rohingya.
- Sweeping Generalization: Broad, unsupported claims targeting the entire Rohingya community, disregarding individual nuances.
- Coded Language: Subtle, subjectively designed phrases implying hostility without overt aggression.
- Fearmongering: Content that exaggerates perceived threats posed by the Rohingya to evoke fear or hostility.

3. In-Group and Out-Group Framing: This category frames the Rohingya as perpetual outsiders, reinforcing an "us versus them" dynamic that fuels discriminatory attitudes and advocates exclusion.

- Dehumanizing: Narratives stripping the Rohingya of humanity, portraying them as inferior

- or undeserving of basic rights.
- **Criminalizing:** Claims associating the Rohingya with illegal activities such as drug trafficking, terrorism, or crime.
- **Stereotyping:** Simplified and overgeneralized portrayals reinforcing biased assumptions about the Rohingya.
- **Othering:** Content emphasizing the Rohingya as outsiders who do not belong, reinforcing divisions between "us" and "them."
- **Discriminatory/Exclusion:** Narratives advocating for unequal treatment, marginalization, or exclusion of the Rohingya from society.
- **Call for Harm/Violence:** Explicit or implicit encouragement of physical harm, violence, or hostility against the Rohingya.

Non-Adversarial Narratives: In contrast to the adversarial narratives, the study also found posts that lacked hostility or aggression toward the Rohingya. These were categorized as non-adversarial narratives. Neutral, which includes posts that lack comments, opinions, representations, or sentiments directed toward the Rohingya community and Positive category comprises posts that exhibit sympathetic, inspirational, or supportive opinions, comments, or representations favoring the Rohingya community. Such narratives were filtered out and excluded from the detailed analysis.

Content: Going beyond the adversarial narrative model, we also categorized the content into two types to better understand it from a Misinformation, Disinformation, and Hate Speech (MDH) perspective. These categories helped us further refine the analysis by distinguishing between the distinct ways that harmful content is propagated:

- **Misinformation/Disinformation:** This includes false, misleading, or exaggerated claims aimed at spreading distrust or reinforcing negative stereotypes and that are already debunked by fact checkers or reported by media outlets.
- **Hate Speech:** Hate content consists of explicit or implicit messages that promote hostility, exclusion, or harm toward the Rohingya.

Actors: Narratives about the Rohingya community are shaped by a complex set of actors, from individuals to organizations, each influencing public perceptions. These actors contribute to both adversarial and non-adversarial content, highlighting the interplay between personal biases, institutional strategies, and organizational agendas. The study identified key actors behind these narratives, including:

- **Individual Accounts:** Private users sharing personal opinions or information.
- **State Actors:** Government's offices and officials shaping narratives.
- **Media Organizations:** News outlets influencing perceptions through reporting.
- **Political Actors:** Politicians and groups advancing strategic or ideological agendas.
- **Religious Actors:** Faith-based groups shaping narratives through religious perspectives.
- **Other :** NGOs, community groups, and other organizations contributing to the discourse.

How we classified hate speech

This study uses a platform-aligned framework to identify hate speech within adversarial narratives. Drawing from the community guidelines of Meta (Facebook) and YouTube, we established a classification system that matches observed content against the stated policies of each platform. Our goal was to ensure that the labeling of hate speech was consistent with

existing enforcement standards, especially in the context of identity-based targeting.

We cross-referenced content types with the policies of both platforms to determine whether a post should be categorized as hate speech. The categories were then divided into two groups: those that clearly qualify as hate speech under both platforms' policies, and those that are context-dependent and may require further qualitative judgment.

Categories that clearly fall under hate speech policies

- **Call for Harm:** Posts that incite violence or serious harm based on identity (e.g., religion, ethnicity) clearly violate both Facebook and YouTube policies.
- **Slurs:** Posts using dehumanizing or degrading language to refer to identity groups.
- **Discriminatory:** Content expressing inferiority, exclusion, or unequal treatment based on protected characteristics.

These categories were considered hate speech if they matched platform criteria—either through explicit wording or implied intent.

- **Fearmongering:** Posts that attempt to incite fear by linking identity groups to violence, disease, or societal harm. Classified as hate speech when the framing is dehumanizing, inciting, or promotes exclusion.
- **Stereotyping:** Posts that apply negative generalizations (e.g., criminality, moral decay) to entire identity groups. Classified as hate speech when such stereotypes imply inferiority or promote hostility.
- **Othering:** *Content that portrays a group as alien or sub-human may qualify if it overlaps with dehumanizing or exclusionary language.*
- **Ridiculing:** *Posts that mock protected groups were assessed for tone, severity, and intent. Satirical or humorous posts were not automatically flagged unless they crossed into overt harm.*

Categories that fall under both Facebook and YouTube hate speech policies

Category	Facebook	YouTube	Hate Speech?
Call for harm	Prohibits content that "calls for violence or serious harm against people based on protected characteristics."	Disallows inciting hatred or violence based on identity (e.g., race, ethnicity, nationality, religion).	✓ Yes
Slurs	Flags use of dehumanizing language or "calls for exclusion" using slurs related to protected groups.	Prohibits slurs intended to "incite hatred" or demean based on identity.	✓ Yes
Discriminatory	Prohibits content expressing "inferiority or exclusion" based on identity.	Bans content that promotes discrimination against protected groups.	✓ Yes

Context-Dependent Categories

Category	Facebook	YouTube	Hate Speech?
Fearmongering	If fearmongering links a group to violence, disease, or harm in a dehumanizing way, it's hate speech.	Same, if it incites hatred or dehumanizes based on identity.	⚠️ Often
Stereotyping	Harmful stereotypes are moderated when they imply inferiority, criminality, or dehumanization.	Prohibited if they promote hatred or discrimination.	⚠️ Often
Othering	Prohibits portraying people as “sub-human”—which overlaps with some forms of othering.	Same, especially if it links identity with inferiority.	⚠️ Sometimes
Ridiculing	Permits humor or satire unless it's clearly targeting protected identities in a degrading way.	May allow ridicule under “borderline content” unless it crosses into hate speech.	⚠️ Sometimes

One key limitation of this methodology is its reliance on the official hate speech definitions provided by platforms like Facebook and YouTube. These definitions are often shaped by corporate risk management and may not fully capture the broader spectrum of harm, particularly in non-English contexts like Bangla or Rohingya. Platform enforcement is also inconsistent, especially in low-resource languages, which means some harmful content may go unmoderated even if it violates stated policies. Despite these shortcomings, we chose to align our classification framework with platform policies because by working within their own rules, we hope to provide evidence that supports better enforcement and more consistent application of standards in protecting vulnerable communities like the Rohingya.

