CENTRE FOR INDEPENDENT JOURNALISM

Advocating Media Freedom and Access to Information

ONLINE NARRATIVES AND DIGITAL THREATS

The Spread of Misinformation, Disinformation, and Hate Speech Against Rohingya Refugees in Malaysia

Supported by: APC Association for Progressive Communications

ONLINE NARRATIVES AND DIGITAL THREATS

The Spread of Misinformation, Disinformation, and Hate Speech Against Rohingya Refugees in Malaysia

CENTRE FOR INDEPENDENT JOURNALISM
Advocating Media Freedom and Access to Information

The Centre for Independent Journalism (CIJ) is a feminist, freedom of expression watchdog and non-profit organisation that aspires for a society that is democratic, just and free, where all peoples will enjoy free media and the freedom to express, seek and impart information

Centre for Independent Journalism (Malaysia)
L13A-1, Menara Sentral Vista,
150, Jalan Sultan Abdul Samad, Brickfields
50470 Kuala Lumpur, Malaysia

Email	cijmalaysia@gmail.com
Website	cijmalaysia.net
Facebook	facebook.com/CIJ.MY
X	twitter.com/CIJ_Malaysia
Instagram	instagram.com/cij_malaysia
YouTube	youtube.com/cijmalaysia

Authors
Suyin Chia
Wathshlah Naidu
Dineshwara Naidu

Project Coordinator
Irfan Naveen Nazrin (Project Coordinator)

Monitors
Norhayati Arham (Lead)
Hannah Hon
Eriq Daniel

Copy Editor:
Lim Jih-Ming

Published in Kuala Lumpur in April 2026

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report investigates the alarming rise and normalisation of online hate speech, misinformation, and disinformation targeting the refugee community in Malaysia, with a particular focus on the Rohingya community. Through detailed monitoring undertaken in 2024 of social media platforms such as Facebook, TikTok, and X (formerly Twitter), it highlights a growing trend of xenophobia and hostility fueled by coordinated narratives, the government, public figures, and vigilante groups. The study identifies key actors, their motives, narratives, and dissemination methods that amplify harmful rhetoric, ultimately endangering the refugee community safety and well-being.

The findings reveal that the framing of refugees as a security, economic, and cultural threat has normalised hostility, resulting in both online abuse and real-world violence. As public sentiments turn increasingly hostile against refugees, social media is becoming a central tool for amplifying this negativity, with bad-faith actors leveraging algorithms to spread divisive and emotionally charged narratives. Social media platforms, prioritising profit over engagement and accuracy, exacerbate the issue by aggressively amplifying disinformation and hate speech.

Additionally, media outlets, both traditional and new media, play a critical role in amplifying hate by sensationalising issues and reinforcing stereotypes. By exploiting public fears and societal tensions, these actors intensify xenophobic rhetoric to drive online engagement, all at the expense of the refugee community safety.

LIST OF ACRONYMS

AI	Artificial Intelligence
CIB	Coordinated Inauthentic Behaviour
CIJ	Centre for Independent Journalism
CMA	Communications and Multimedia Act
CSO	Civil Society Organisation
ICCPR	International Covenant on Civil and Political Rights
KDN	Kementerian Dalam Negeri (Ministry of Home Affairs/MOHA)
MDH	Misinformation, Disinformation and Hate Speech
PATI	Pendatang Asing Tanpa Izin (illegal migrant)
RMP	Royal Malaysia Police
UGC	User-Generated Comments
UMNO	United Malays National Organisation
UNHCR	United Nations High Commissioner for Refugees

01 INTRODUCTION

Malaysia has been hosting refugees on its territory since the 1970s when Vietnamese refugees began to arrive by boat in Malaysia and other countries in the region. From 1975 to 1996, Malaysia hosted some 240,000 Vietnamese refugees. From the 1970s to the 1990s, Malaysia also hosted over 50,000 Filipino Muslims, several thousand Cambodian Chams, and several hundred Bosnian refugees. Today, Malaysia hosts some 192,200 refugees and asylum-seekers from over 50 countries. The majority are individuals from conflict-affected areas or those fleeing persecution in Myanmar. This includes some 111,700 ethnic Rohingya refugees, making it one of the largest Rohingya diaspora communities outside of South Asia.[1] Although Malaysia is not a signatory to the 1951 Refugee Convention, Malaysia has historically accorded relative safety to refugees in the country, including the Rohingyas, while officially classifying them as undocumented migrants.

This classification complicates their legal status and restricts access to critical services and protection, including education, healthcare, and formal employment. As a result, many live in precarious conditions, vulnerable to exploitation, extortion, arrests, and detention.[2]

The influx of Rohingya refugees arriving in Malaysia continues to rise. In 2023 alone, more than 4,400 refugees made the dangerous journey by sea—a 20 per cent increase from previous years.[3] Many of these arrivals come from Cox's Bazar in Bangladesh, where over 1 million Rohingya refugees have lived in overcrowded camps for years. As conditions in the camps deteriorate and the civil war in Myanmar escalates, experts warn that more Rohingya refugees may attempt to flee to Malaysia, India, and Indonesia in search of safety and stability.

Malaysia, once known for its welcoming stance towards Rohingya refugees and for largely overlooking their technically illegal presence in the country, has seen public sentiment shift in recent years. Bipartisan support and public demonstrations previously reflected empathy and solidarity with the plight of the Rohingya. In 2016, then Prime Minister Najib Razak joined a protest rally, declaring, "We must defend them [Rohingya] not just because they are of the same faith but they are humans, their lives have values."[4]

In 2020, public sentiments towards Rohingya refugees in Malaysia underwent a significant transformation, with widespread hostility and hate speech generally targeting Rohingya refugees. This shift in narrative was triggered primarily by the arrival of Rohingya refugees via boats in April that year and reported incidents of boat pushbacks by authorities. Several

---

1 UNHCR. (2022, August 9). Figures at a glance in Malaysia. UNHCR. https://www.unhcr.org/my/what-we- do/figures-glance-malaysia

2 *Rohingya are at risk "wherever we go, whether Myanmar, Thailand, or Malaysia."* (2024). Doctors without Borders - USA. https://www.doctorswithoutborders.org/latest/rohingya-are-risk-wherever- we-go-whether-myanmar-thailand-or-malaysia

3 Ibid.

4 *Malaysian PM leads protest against "genocide" of Rohingya*. (2016, December 4). AP News. https://apnews.com/9701bb6b50ba46b3aff5ff071957ec73/malaysian-pm-leads-protest-against- genocide-rohingya

other factors also contributed to this rise in anti-Rohingya sentiment, including Malaysia's heightened economic and political instability due to an unprecedented take-over of the previous government in February 2020, against the backdrop of a global pandemic and nationwide lockdown enforced through Movement Control Orders (MCO), which provided a captive online audience to share and comment on negative MDH content.

During the pandemic lockdown, the perception of migrants and refugees, including Rohingyas, shifted from being the mostly 'invisible other' to a 'threat' to the health and economic well- being of Malaysians. Populist sentiments framed Rohingya as 'outsiders', coinciding with reinforced anti-Rohingya messages from various sources. An enforcement agency, for example, posted a controversial poster on Facebook declaring "Ethnic Rohingya migrants, your arrival is unwelcome".[5] This period also saw a heightened crackdown against human rights defenders, including refugee rights advocate Heidy Quah, who was critical of government actions.[6] Posts announcing enforcement actions, particularly those referencing undocumented migration, frequently attract extensive user-generated commentary. Analysis indicates that within these comment sections, hostile or discriminatory speech can emerge and escalate rapidly, often extending beyond the scope of the original administrative message. If left unmoderated, such rhetoric risks normalizing dehumanization, encouraging harassment, and, in some cases, contributing to real-world harms, including intimidation or violence. Accordingly, pr oactive measures to curb and manage these comments are essential to prevent escalation and to mitigate potential adverse impacts beyond the online space.

Since 2021, calls to protect national interests have amplified discriminatory rhetoric towards undocumented migrants, refugees, and asylum seekers, especially the Rohingya. The continuous arrival of refugees, coupled with media reports linking Rohingyas to public order disturbances,[7] has fuelled increasingly hostile sentiments. Certain quarters have come to perceive them as a socioeconomic and security threat, pressuring the government to halt the influx. Local complaints have also risen, accusing the Rohingya of operating businesses without licences and using government land without permission.[8] Some have gone so far as to demand the expulsion of humanitarian organisations, particularly the United Nations High Commissioner for Refugees (UNHCR), from the country.

The rise in hostility has not remained confined to hate speech and rhetoric. Vigilante attacks against Rohingya refugees have escalated in urban areas such as Selayang and Penang, where large Rohingya communities reside. For example, in April 2023, a community of over 50 Rohingya refugees, including children and senior citizens, living in rural Penang were

---

5 *Malaysia's Anti-Rohingya Refugee Poster Angers Rights Group.* (2021, June 11). Benar News. https://www.benarnews.org/english/news/malaysian/malaysia-anti-rohingya-immigration- 06112021161832.html

6 https://www.malaymail.com/news/malaysia/2022/04/25/refuge-activist-heidy-quah-given-discharge- not-amounting-to-acquittal-for-i/2055559

7 Rahim, N.F.A., & NOH, M.F. (2024, January 21). *"Colonies" of Rohingya taking over different locations nationwide.* NST Online. https://www.nst.com.my/news/nation/2024/01/1003822/colonies-rohingya-taking- over-different-locations-nationwide

8 Sukhani, P. (2020, July 10). *The Shifting Politics of Rohingya Refugees in Malaysia*. Thediplomat.com. https://thediplomat.com/2020/07/the-shifting-politics-of-rohingya-refugees-in-malaysia/

driven from their homes by local vigilantes "without provocation".[9] In August 2024, the CIJ team identified plans by members of the vigilante group SURPLUS to confront the UNHCR Malaysia office with claims "to hold the organisation accountable". The team promptly provided an early warning to UNHCR staff, enabling them to prepare for the encounter.

At the same time, regional dynamics also play a significant role in shaping public sentiment and discourse about refugees in Malaysia. For example, Aceh, Indonesia, which faces similar challenges with refugee arrivals, has shifted its public opinion from empathy to hostility towards the Rohingya. In December 2023, hundreds of university students in Aceh stormed a temporary shelter housing Rohingya refugees, demanding their immediate deportation. This incident followed Aceh's largest influx of Rohingya refugees in eight years, with more than 1,500 arrivals in a short period. UNHCR attributed the attack to a coordinated online campaign of misinformation and hate speech.[10] The interplay between regional developments and local attitudes underscores how misinformation campaigns and hateful narratives across borders contribute to a growing climate of hostility towards Rohingya refugees.

The rise of xenophobic online campaigns and even vigilante attacks against the Rohingya community underscores the dangerous trajectory of this growing hostility. In the era of digital connectivity, Malaysia's social media landscape has become central in shaping public perceptions. Online platforms such as Facebook, X (formerly Twitter), and TikTok have become breeding grounds for MDH, worsening the already fragile situation faced by the Rohingya.[11] Online campaigns targeting the Rohingya exploit broader societal tensions, framing refugees as threats to national security and public order. These harmful narratives spread rapidly, normalising hostility and escalating calls for discriminatory policies and government action against refugees.

Further complicating the digital space, cybertroopers—actors employed to manipulate public opinion—actively spread negative narratives and hate speech against the Rohingya for profit and political clout. By exploiting algorithms on social media platforms, these actors ensure that harmful content gains visibility, influencing public sentiments and pressuring the authorities to adopt anti-refugee policies.

This report will map the patterns and severity of MDH targeting Rohingya refugees in Malaysia. It will identify the actors and networks involved in spreading xenophobic narratives and assess the implications for social cohesion and human rights in the country. Through this analysis, the report aims to provide insight into the growing hostility towards the Rohingya and offer recommendations for mitigating hate speech and protecting vulnerable communities.

---

9 Penang Rohingya villagers forced from homes ahead of Raya. (2023, April 19). MalaysiaNow. https://www.malaysianow.com/news/2023/04/19/penang-rohingya-villagers-forced-from-homes-a-week-before- raya

10 *UNHCR disturbed over mob attack and forced eviction of refugees in Aceh, Indonesia.* (2023, December 11). UNHCR Asia Pacific. https://www.unhcr.org/asia/news/press-releases/unhcr-disturbed-over-mob-attack-and- forced-eviction-refugees-aceh-indonesia

11 Zainul, H. (2020, June 24). *Campaign of Hate? Fake News and Anti-Refugee Rhetoric in Malaysia.* ISIS. https://www.isis.org.my/2020/06/24/campaign-of-hate-fake-news-and-anti-refugee-rhetoric-in-malaysia

1.1 Legal Framework Against Hate Speech

Addressing hate speech requires balancing freedom of expression with protection against discrimination, hostility, and violence. However, no universally accepted definition of hate speech exists.

Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR)[12] highlight the obligations of States regarding freedom of expression and hate speech. Article 20(2) mandates that States prohibit national, racial, or religious hatred that incites discrimination, hostility, or violence. At the same time, Article 19(3) allows restrictions on freedom of expression only under exceptional circumstances, ensuring such limitations meet the following three-part test:

i. **Legality:** Restrictions must be provided for by law
ii. **Legitimacy:** Restrictions must pursue a legitimate aim, such as protecting the rights and reputation of others
iii. **Necessity and Proportionality:** Restrictions must be necessary in a democratic society and proportionate to the aim pursued

In addition to these requirements, the Rabat Plan of Action (2012)[13] sets a high threshold for limiting speech by adding a six-part test to the ICCPR framework. This threshold ensures that only serious cases of incitement to hatred warrant restrictions. The Rabat test assesses:

i. **Context:** The social, political, and economic factors of the statement
ii. **Status of the speaker:** The influence and authority of the individual
iii. **Intent:** Whether the speech aims to incite violence against a target individual or group
iv. **Content and form:** The language and framing of the speech
v. **Extent of dissemination:** How widely the speech has spread
vi. **Likelihood and imminence of harm:** Whether the speech is likely to provoke immediate harm

Malaysia's legal framework offers some tools for regulating harmful speech, but often falls short of international standards established by the ICCPR and the Rabat Plan of Action. Freedom of speech, while constitutionally protected under Article 10 of the Federal Constitution, is subject to limitations that address expressions relating to race, religion, royalty, public morality and national security. However, Malaysia lacks a specific legal definition for hate speech, making it challenging to establish a clear threshold for what constitutes punishable speech, particularly when directed toward non-citizens and

12 United Nations. (1966, December 16). *International Covenant on Civil and Political Rights*. OHCHR; United Nations. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and- political-rights

13 *OHCHR. The Rabat Plan of Action*. OHCHR. https://www.ohchr.org/en/documents/outcome- documents/rabat-plan-action

marginalised groups.

Several existing laws address forms of speech that threaten public order or social harmony. The Sedition Act 1948[14] criminalises expressions that incite hatred or contempt against the government or any ethnic group. Section 233 of the Communications and Multimedia Act (CMA) 1998[15] also penalises offensive or menacing content transmitted online, while Sections 504 and 505 of the Penal Code[16] target speech that provokes public outrage or incites violence.

Despite these provisions, non-citizens, including refugees and asylum seekers, remain vulnerable. Access to these laws and justice is often limited, given the vulnerability of Rohingya with limited resources and often living in fear if harassment and intimidation due to their 'undocumented status'. Even when online hate speech targets vulnerable minorities, enforcement remains inconsistent. Government efforts tend to prioritise curbing speech deemed politically sensitive, rather than addressing xenophobic content targeting communities at risk. As a result, hate speech against Rohingya often persists without significant legal consequences, further marginalising this already vulnerable group.

1.2 Social Media Platform Addressing Hate Speech

Social media platforms play a critical role in regulating hate speech, defining it within their terms of service or community standards. These platforms, including Facebook, X (formerly Twitter), and TikTok, have established guidelines to address harmful content, often targeting expressions that incite violence, discrimination, or hostility based on protected characteristics. However, the enforcement of these standards in Malaysia remains inconsistent, especially regarding hate speech directed towards communities at risk like the Rohingya.

In May 2016, Facebook, Twitter, and YouTube signed a code of conduct with the European Commission to counter illegal hate speech online.[17] In January 2025, the European Union revised this code to align it with the EU Digital Services Act 2022.[18] While this code sets a precedent for online regulation, its effectiveness in non-European contexts—such as Malaysia —remains limited, given the diversity and nuances of the various languages used in the country, making it difficult for automated detection and moderation.

1.3 Platform-specific Hate Speech Policies

1. **Facebook:** Facebook's Community Standards[19] define hate speech as a direct

---

14 Federal Legislation. "Act 15 Sedition Act 1948." The Commisioner of Law Revision, Malaysia, 2006. https://lom.agc.gov.my/act-detail.php?act=15&lang=BI&date=01-01-2006&timeline.

15 Malaysian Communications and Multimedia Commission. "Communications and Multimedia Act 1998 (Act 588), Incorporating Latest Amendment – Act A1220/2004." Malaysia, 2006. https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/Act588bi_3.pdf.

16 The Commisioner of Law Revision. "Penal Code (Act 574), As at 18 December 2018." 2018.

17 European Commission. (2016, May) *Code of Conduct on Countering Illegal Hate Speech Online*. https://ec.europa.eu/newsroom/just/items/54300/en

18 European Commission. (2025, January) Commission welcomes the integration of the revised Code of conduct on countering illegal hate speech online into the Digital Services Act.

19 Facebook Policies & Guidelines / Facebook for Creators. (2022). https://creators.facebook.com/stay- safe/policies-and-guidelines

attack on individuals or groups based on protected characteristics, including race, ethnicity, national origin, disability, religion, caste, sexual orientation, gender identity, sex, or serious disease.[20]

2. **X (formerly Twitter):** X's Hateful Conduct Policy[21] sanctions users for content that makes violent threats against individuals or groups; incites or encourages harm towards others; references mass violence or historical events targeting protected groups; incites discrimination or hostility towards groups with protected characteristics; repeats racist or sexist tropes or uses slurs and epithets non consensually; and displays or promotes hateful imagery. Although these platforms have established policies against hate speech, enforcement in Malaysia remains inconsistent. While reporting mechanisms exist, civil society groups often encounter delays or inaction when requesting the removal of offensive content.
3. TikTok: TikTok prohibits "any hateful behavior, hate speech, or promotion of hateful ideologies" through its Community Principles.[22] It defines protected attributes as caste, ethnicity, national origin, immigration status, race, religion, gender identity, sexual orientation, and disability. The platform also monitors the spread of hate ideologies, particularly racial supremacy, anti-LGBTQIA+ rhetoric, misogyny, and antisemitism.

Although these platforms have established policies against hate speech, enforcement in Malaysia remains inconsistent. While reporting mechanisms exist, civil society groups often encounter delays or inaction when requesting the removal of offensive content. The increasing reliance on automated moderation tools further compounds the issue. AI moderation, which, while efficient in handling large volumes of content, lacks the ability to interpret context effectively. These systems operate in a binary manner and cannot efficiently detect nuances of local slang, coded language, and cultural context embedded in hate speech—especially in countries like Malaysia, where large, diverse datasets to train AI for linguistic and cultural accuracy are often lacking. As a result, AI moderation frequently generates inaccurate and inconsistent enforcement outcomes, allowing harmful content to persist while sometimes mistakenly taking down non-harmful speech.

This trend is set to deepen as social media platforms continue shifting toward AI-driven content moderation. In October 2024, TikTok announced mass layoffs affecting its global workforce, including several hundred employees in Malaysia, as part of its shift towards greater reliance on AI content moderation, reflecting a broader industry trend of social media companies automating the regulation of online expression.[23]

---

20 Meta, the parent company of Facebook, made sweeping changes to its community standards in January 2025, eliminating many clauses that previously banned specific derogatory statements about protected groups. These changes raise concerns about broad repercussions, particularly by allowing targeted attacks on women, transgenders, LGBTQIA+ people, and immigrants.

21 X's policy on hateful conduct / X Help. (2023, April). Help.x.com. https://help.x.com/en/rules-and- policies/hateful-conduct-policy

22 Community Principles. (2023, March 3). TikTok. https://www.tiktok.com/community- guidelines/en/community-principles

23 Latiff, R. (October, 2024). *ByteDance's TikTok cuts hundreds of jobs in shift towards AI content moderation*. Reuters. https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai- moderation-sources-say-2024-10-11/

The lack of coordination between technology companies, Malaysian authorities, and civil society organisations has resulted in fragmented and ineffective efforts to combat online hate, highlighting the urgent need for more effective measures to refine content moderation strategies and address enforcement gaps.

02 MONITORING FRAMEWORK

This report aims to identify patterns of xenophobic hate speech, misinformation, and disinformation targeting Rohingya refugees in Malaysia across social media platforms. The monitoring period spans 1st June to 31st August 2024, focusing on posts containing xenophobic narratives, misinformation and disinformation, and hate speech against refugees, with particular attention to the Rohingya. The analysis explores the prevalence, underlying motivations, and dissemination methods of these narratives through language analysis, sentiment detection, and contextual interpretation.

2.1 Scope of Monitoring

This project monitored Facebook, X (formerly Twitter), and TikTok, analysing online discourse surrounding Rohingya refugees and migrant issues in Malaysia. The report focused on posts targeting these communities, examining the patterns in the content origin and sources, dissemination netowrks, modes of amplification, and the recurring narratives promoted by the following categories of key actors: politicians and political parties; media outlets; government agencies; and key opinion leaders (KOLs) and influencers. These actors were chosen on the basis of the pattern observed during CIJ's Social Media Monitoring of the 15th General Elections[24] initiative, which noted an escalation of the severity of harmful and hate speech against refugees and migrants by these actors. The analysis considered how harmful biases and toxic rhetoric emerge and circulate across digital spaces, without attributing direct intent to specific institutions or individuals.

The selected platforms were chosen based on their widespread usage and influence in Malaysia's digital landscape. Facebook, with 22.34 million Malaysian users, remains the most dominant platform. X (formerly Twitter), though smaller with 5.71 million users, plays a crucial role in shaping political and social narrative as the second-largest text-based platform, particularly through real-time discussions and viral trends. Meanwhile, TikTok, as the fastest-growing video platform in Malaysia with an advertising reach of 27 million users, is the fastest-growing platform offering a highly engaging space for content consumption and influence.[25] Notably, TikTok was a key battleground for political messaging and a hotspot for racially charged content during Malaysia's 15th General Election,[26] underscoring its power to drive political engagement and discourse.

---

24 Hamzah, M., Naidu, W., Lee, S. F., Naidu, D., & Balachandar, D. (2023). Social Media Monitoring of Malaysia's 15th General elections, https://cijmalaysia.net/social-media-monitoring-of-malaysias-15th-general-elections/

25 Kemp, S. (2024, February 23). *Digital 2024: Malaysia. DataReportal*. https://datareportal.com/reports/digital-2024-malaysia

26 Dzafri, D. (2022, November 22). *Post GE15, TikTok has become a breeding ground for racially charged content. What is the platform doing about it?* Malay Mail. https://www.malaymail.com/news/malaysia/2022/11/22/post-ge15-tiktok-has-become-a-breeding-ground-for- racially-charged-content-what-is-the-platform-doing-about-it/41295

Although not included in this research, other widely used platforms also contribute to Malaysia's digital ecosystem. YouTube remains a dominant force with 24.1 million users, serving as a key platform for long-form content. Instagram, with 15.7 million users, is a popular space for visual storytelling. LinkedIn, while more niche as a professional networking platform, has a growing presence with 7.8 million members.

The report identifies disinformation, misinformation, and divisive narratives that contribute to amplifying hate speech and hostility toward the Rohingya community. The analysis focuses on the following themes:

- **Security Threat:** Portraying refugees as risks to national security
- **Labour or Economic Threat:** Framing refugees as competitors in the job market
- **Crime and Public Disorder:** Associating refugees with rising crime and social disturbances
- **Race and Religion:** Promoting racial and religious intolerance against refugees
- **Pro-zionist Narratives:** Using divisive political rhetoric to sow discord
- **Misogynistic and Sexist Narratives:** Spreading gender-based hate towards refugees
- **Discrediting UNHCR:** Undermining humanitarian organisation's efforts and credibility

The project also tracks disinformation campaigns that amplify hate speech and spread hostility towards Rohingya. It further detects AI-generated content and traces the presence of Coordinated Inauthentic Behavior (CIB) efforts.

2.2 Defining Hate Speech for the Monitoring Project

As there is no universally agreed definition of hate speech or levels of hate speech, including in Malaysia, CIJ adapted its own levels. The following hate speech severity framework was first developed and tested for the CIJ's Social Media Monitoring of Malaysia's 15th General Elections report published in March 2023.[27]

The hate speech severity framework establishes the parameters of hate speech and categorises speech on a spectrum of four severity levels with the corresponding characteristics:

Level 1: Disagreements or non-offensive language

Level 2: Offensive or discriminatory language

Level 3: Dehumanising or hostile language

Level 4: Causing incitement or calls for violence

---

27 Hamzah, M., Naidu, W., Lee, S. F., Naidu, D., & Balachandar, D. (2023). *Social Media Monitoring ofMalaysia's 15th General elections [Review of Social Media Monitoring of Malaysia's 15th General elections].* In G. Venkiteswaran, Z. Nain, S. S. Ngo, M. Ahmad, L. K. Wang, K. T. Lee, L. Lai Chee Ching, J. A. Surin, &
I. A. Ismail (Eds.), cijmalaysia.net. Centre for Independent Journalism Malaysia.

The hate speech framework further adapted the threshold test in the Rabat Plan of Action using the 3M framework in the following table:

Table 1: Threshold test for hate speech severity

MESSENGER	1. Speaker	- Position/Status (degreee of influence) in the society.
	2. Intent	- Intent to incite hatred of inflict harm.
MESSAGE	3. Content	- Locating the speech within social, religious and political context, and the power dynamics at the time the speech was made and disseminated. This may include past historical context, policies, or social norms.
	4. Content & Form	- Degree to which speech was provocative and direct. - Form, style and nature of arguments used in the speech.
AUDIENCE	5. Audience	- Reach of speech, its public nature, magnitude and audience size. - Frequency, quantity and extent of communications. - Whether the statement is circulated in a restricted environment or widely accessible to the public. - Whether the audience had the means to act on the incitement.
	6. Medium	- Means of messaging and amplification (one platform or multiple or cross-platforms). - Use of inauthentic accounts (bots, cybertroopers).

2.3 Defining Disinformation, Misinformation, and Malinformation

This monitoring report classifies different forms of disinformation along a spectrum based on their intent to mislead and the extent to which they distort facts and reality. CIJ established an information disorder systems framework using this dimension of harm and falseness to distinguish between misinformation, disinformation, and malinformation to effectively monitor harmful narratives targeting Rohingya refugees. Each type represents a distinct aspect of information disorder that requires tailored responses.

Key Characteristics of Disinformation, Misinformation, and Malinformation[28]

1. **Disinformation** refers to false or misleading information intentionally created

28 Reppell, L., & Shein, E. (2019). Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions [Review of Disinformation Campaigns and Hate Speech: Exploring the Relationship and Programming Interventions]. In ifes.org. International Foundation for Electoral Systems. https://www.ifes.org/sites/default/files/migrate/2019_ifes_disinformation_campaigns_and_hate_speech_briefing_paper.pdf

and spread to deceive, cause harm, or benefit the perpetrator. The harm may target individuals, communities, institutions, or societal processes (such as refugee protections, elections, etc.) Disinformation is motivated by three main factors: for profit, to gain political influence, or to cause trouble for the sake of it.

2. **Misinformation** refers to inaccurate information shared without intent to harm. When disinformation is shared, it often turns into misinformation. For example, when individuals unknowingly share false content without thoroughly checking if the information is accurate or misleading.
3. **Malinformation** involves accurate information shared with malicious intent to inflict harm, including to embarrass, discredit, or damage reputations, by leaking or weaponizing private or confidential information. Examples of malinformation include phishing and doxing.

Table 2: Parameters of information disorder spectrum

Content Type	Description	Intent	Classification
Satire or Parody	Mimics real news in a humorous or mocking way; not intended to harm but can mislead if taken seriously.	No harmful intent, but may deceive	Recognised as a legitimate form of expression. However, often classified as a
Misleading Content	Selective use or framing of information to distort meaning or perception of an issue or individual.	Manipulates facts to mislead	Misinformation
False Connection	When headlines, visuals, or captions do not align with the actual content.	Misleads through mismatched framing	Misinformation
Manipulated Content	Entirely false content deliberately created to deceive and cause harm.	Creates deception through distortion	Disinformation
Fabricated Content	Entirely false content deliberately created to deceive and cause harm.	Fully intended to deceive and harm	Disinformation

2.4 Defining Coordinated Inauthentic Behavior

CIJ defines coordinated inauthentic behavior (CIB) as actions by individuals or groups that collaborate to deceive others about their identities and activities. This behaviour is identified not by the content they share but by the deceptive methods they use. These actors manipulate online engagement through the use of algorithms, bots designed to mimic human behaviour, or 'cybertroopers' who create fake profiles, posts, and likes. Their goal is to amplify specific narratives or redirect public attention towards targeted individuals, groups, or topics.

The following rules were used to determine if content was generated by an inauthentic account:

a. If an account was recently activated
b. If an account had limited followers
c. If an account had similar types of messages and responses
d. If an account was following similar types of accounts and reposting similar content
e. The speed, frequency, and focus of dissemination and response

03 METHODOLOGY

Monitoring Approach and Methodology

The monitoring process consists of two phases, utilising Zanroo, an automated social media monitoring tool, in conjunction with human review for more detailed analysis. This enabled the scanning and processing of thousands of online contents across various platforms, enabling the 'scraping' of data that matched the predefined keywords and accounts identified by CIJ monitors.

A total of 57 keywords were generated through consultation with affected communities and civil society organisations (CSOs) working on refugee issues. The keyword list was continuously updated and adapted during the monitoring period to reflect the evolving narratives, trends, and subtle language shifts in real time. These keywords were used for both manual and automated searches to identify relevant text posts, videos, graphics, accounts, hashtags, and comments. The keywords utilize a Boolean search system to track keywords related to xenophobic rhetoric and misinformation targeting refugees in Malaysia. By combining keywords with Boolean operators (AND, OR, NOT), the system refines search queries to capture relevant discourse while filtering out unrelated content, ensuring precise data collection across social media platforms.

A team of CIJ monitors, trained in reviewing and categorizing content, tagged the dataset according to the keywords, the information disorder system framework, and the severity levels of hate speech.

Additionally, the CIJ monitoring team created two social media accounts on TikTok to test and leverage the platform's algorithm for curating targeted content. This was done by engaging with posts containing xenophobic language, hashtags related to refugee issues, and content from key opinion leaders and influencers known for anti-foreigner and anti-refugee rhetoric. By following, bookmarking, and liking strategically on such posts, the CIJ accounts were able to mimic user patterns, effectively triggering the algorithm to prioritise and surface similar content. This approach allowed CIJ to capture a broader and more nuanced 358 dataset that reflects real-time trends.

This process resulted in the extraction of 19,540 posts through an initial automated scraping. This figure includes 358 posts on TikTok derived through CIJ manual monitoring.

After eliminating duplicates, irrelevant posts, and posts that fell outside the scope of the monitoring framework by human monitors, the dataset was refined to focus on 14,176 posts for deeper analysis.

Phase 1: Hate Speech and Divisive Narratives

i. Review and Tagging: Monitors identified and tagged hate speech levels based on

content posted by key actors, including politicians, media outlets, agencies, and influencers, to assess the intensity of language and its potential to incite harm.

ii. **Pattern and Amplification Analysis:** Monitors analysed how these actors amplified harmful content across networks and identified key amplification methods used to spread hate speech and harmful narratives.

Phase 2: Discrediting Campaigns

i. **Identification of Campaigns**: The dataset was organised into thematic buckets to identify and trace trending narratives and discrediting campaigns. These were analysed based on their alignment with various types of disinformation. This analysis aimed to map a spectrum of disinformation campaigns, focusing on their intent, accuracy, and the harm they caused by spreading hostility towards Rohingya refugees.

ii. **Behavioural Analysis:** This phase involved assessing how coordinated efforts, including AI-generated content and CIB operations, shaped narratives and fueled negative sentiment against the Rohingya.

Finally, CIJ implemented targeted reporting for content categorised as especially harmful, particularly posts falling under Level 3 and Level 4 severity on the hate speech framework. These posts were flagged and reported to social media platforms. This proactive approach aimed to mitigate the spread of the most damaging narratives and reduce the immediate risks posed to the safety and well-being of the Rohingya community.

Supporting Data Sources

The findings are validated by cross-referencing a comprehensive array of supporting data sources, including public records, government and law enforcement documents, newspaper articles, and scholarly publications.

Community-Based Approach

The report is underpinned by a community-based approach by including Rohingya refugees throughout the project to understand how hate speech and misinformation impact them directly.

Limitations

This report is subject to several key limitations that warrant caution when interpreting its findings.

1. Firstly, the dataset excludes content in Chinese and Indian languages due to expertise and resource constraints. While this decision was made to streamline data collection and analysis within the available budget and time frame, it limits the report's ability to fully capture perspectives and trends among communities where these languages are predominantly spoken, and as such, the insights presented may not reflect the full diversity of the Malaysian population.
2. Second, while online private messaging applications such as WhatsApp and

Telegram, are significant platforms where misinformation can spread rapidly, due to privacy concerns and restricted access to these closed networks, the scope of monitoring was limited to publicly available online content. This restriction means that potential misinformation or disinformation circulating in private channels cannot be captured, and therefore, our findings may underestimate the overall scale and dynamics of misinformation across Malaysian communication platforms.

3. Third, while efforts were made to select a comprehensive set of keywords, this approach may not fully capture the complexity of how hateful speech is communicated. The dataset may miss out on content shared using coded language, euphemisms, emojis, and local slang, which often change rapidly in response to social and cultural shifts.

04 SOCIAL MEDIA PLATFORMS

As of January 2024, approximately 83.1 per cent of the Malaysia population are active social media users.[29] With the trend of declining consumption of mainstream traditional media sources such as television news and print media, social media has emerged as an increasingly primary source for news consumption. Platforms such as Facebook, WhatsApp, YouTube, and TikTok dominate the new media landscape, acting as a convenient and accessible news source for many Malaysians.

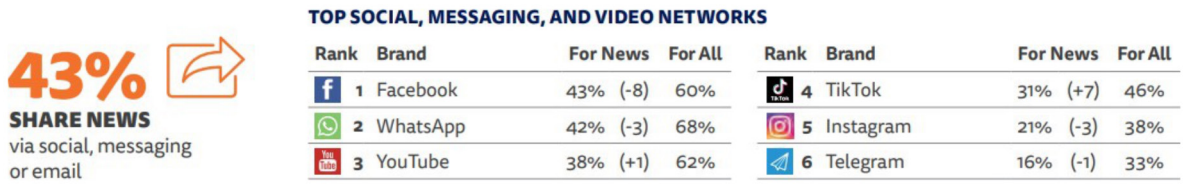


Figure 1: Credit: Reuters Institute

In response to the growing dominance of social media platforms, many traditional news outlets in Malaysia have expanded their digital presence to online formats. Several major publications and news aggregators operate as digital-only platforms, providing news through apps and social media channels.

Social media platforms have made it easier for users to engage with and share news through cross-platform functions. For example, Facebook, TikTok, and X enable seamless content sharing, allowing users to distribute articles, videos, and posts across multiple platforms and boosting visibility across different networks with a single tap. This cross-platform functionality accelerates the viral spread of information, making both reliable news and disinformation more accessible to wider audiences at unprecedented speeds. This viral spread increases news reach and makes it harder to contain once it gains momentum.

29 Statista. (2024). Malaysia: social media penetration 2024.https://www.statista.com/statistics/883712/malaysia-social-media-penetration/

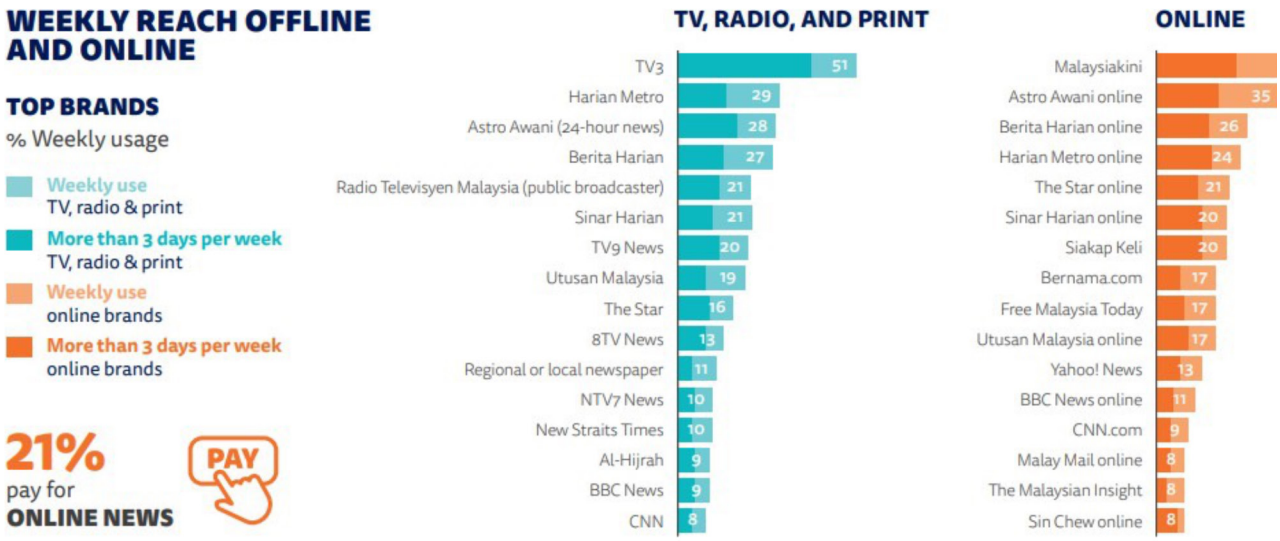


Figure 2: Credit: Reuters Institute.

4.1 How Algorithms Prioritise Engagement over Accuracy and Amplify Misinformation

Social media platforms operate on engagement-driven business models where user attention translates directly into revenue.[30] Their algorithms prioritise content that generates high engagement, such as likes, shares, and comments, to keep users on the platform longer, relatedly, to increase advertising revenue. However, this focus on engagement comes at the cost of accuracy and quality.[31] The systematic prioritisation of user engagement over factual accuracy has unintended consequences, as polarising and emotionally charged content often outperforms factual content in terms of user interaction.

This prioritisation creates environments where xenophobic narratives, misinformation and disinformation campaigns, and hate speech thrive. Below are key ways in which algorithms contribute to this dynamic:[32]

1. **Emotionally Triggering Content:** Algorithms reward content that evokes strong emotions such as fear, outrage, or anger, as it keeps users engaged for longer. Disinformation campaigns exploit these emotions by mobilising propaganda vilifying refugees and communities at risk to generate higher user interaction, making it more likely to appear in user feeds to gain traction and fuel public hostility.
2. **Viral Feedback Loops:** Algorithms promote popular posts and encourage cross-platform sharing (for example, from Facebook to TikTok and X). As a post gains

30 Narayanan, A. (2023, March 9). Understanding social media recommendation algorithms. Knight First Amendment Institute. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

31 Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-34203-2

32 Metzler, H., & Garcia, D. (2023). *Social Drivers and Algorithmic Mechanisms on Digital Media.* Perspectives on Psychological Science, 19(5), 735-748. https://doi.org/10.1177/17456916231185057

more attention, algorithms on multiple platforms reinforce each other, creating a feedback loop that makes it difficult to contain or correct the false information.

3. **Precision Targeting, Filter Bubbles, and Echo Chambers:** Algorithms cluster users based on their interests, interactions, and behaviour patterns, creating niche communities known as filter bubbles. This allows advertisers and content creators to target users with extreme precision and increase the likelihood that users will encounter content aligned with their interests. These communities reinforce biases by encouraging users to spend more time with like-minded communities, creating echo chambers that make users more susceptible to misinformation and divisive narratives.
4. **Coordinated Inauthentic Behaviours (CIBs):** Inauthentic actors, including cybertroopers and AI-generated accounts, use algorithms to coordinate campaigns, posting identical or similar content across multiple accounts at the same time to encourage the viral spread of certain information. This approach mimics popular engagement on these platforms by manipulating the algorithm to detect this surge of activity as a trending topic, amplifying the content further and, therefore, mainstreaming certain narratives.

05 KEY FINDINGS

5.1 Overall Highlights

Table 3: Severity of Content August-September 2024

Severity Level	Unique Content
Level 1: Disagreement/non-offensive	11,889
Level 2: Offensive/discriminatory	2,014
Level 3: Dehumanising/hostile	244
Level 4: Incitement/call for violence	29
TOTAL UNIQUE CONTENT	**14,176**

The report's dataset contained 14,176 unique posts across Facebook, TikTok and X (formerly Twitter). While most posts fall into Level 1, the presence of higher-severity content (Levels 3 and 4) signals a troubling normalisation of extreme hate speech, with incitement to violence constituting a small but not insignificant subset.

5.2 Social Media Platforms By Severity Level:

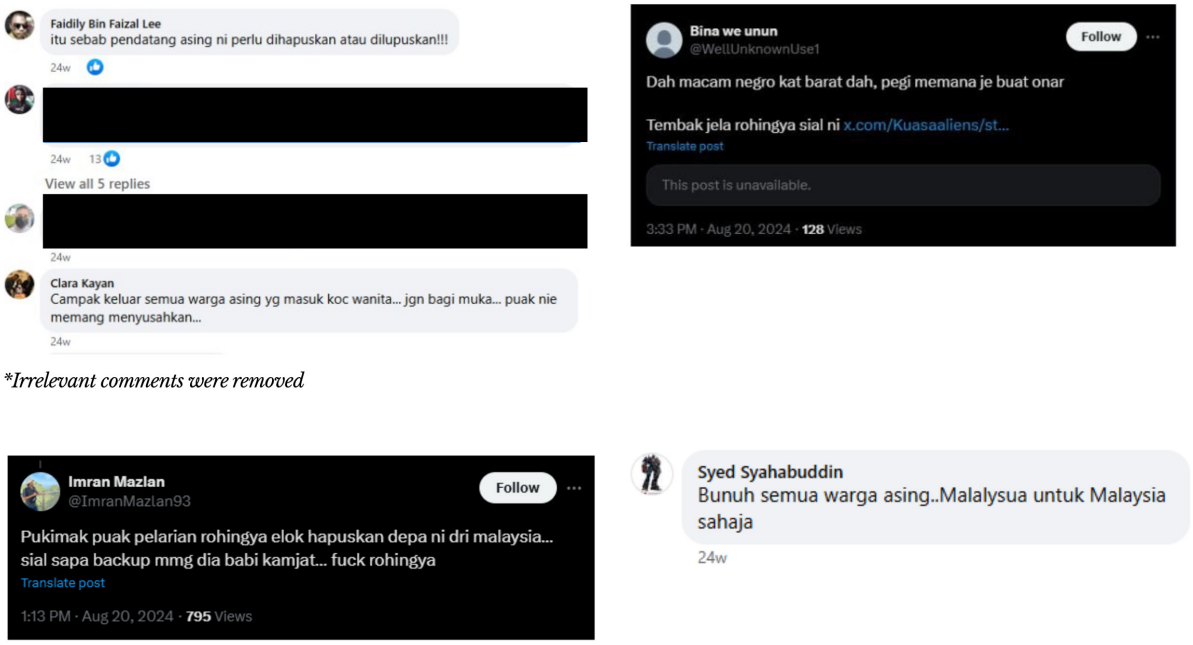
Table 4: Social Media Platforms By Severity Level

Platform	Total Messages	Level 1	Level 2	Level 3	Level 4
Facebook	7,730	6,607	1,036	129	7
TikTok	714	693	20	1	
X (Twitter)	5,751	4,865	832	76	1
TOTAL	**14,195**	**12,165**	**1,888**	**206**	**8**

Facebook accounts for the highest number of hate speech posts across all severity levels among the social media platforms analysed, while X (formerly Twitter) similarly reflects a significant level of activity. The majority of engagement occurs in the comments sections, where users coalesce to express opinions in response to posts, particularly those made by official entities. The most severe cases in Level 3 and Level 4 severity, including direct incitement to violence and hate, were predominantly user-generated comments found under posts by enforcement agencies. These posts serve as the initial trigger for discussions and it is

the subsequent user-generated comments that often escalate into hate speech. Furthermore, while most comments are classified under Level 1, the transition to higher severity levels demonstrates how discussions can escalate quickly in environments with minimal moderation.

TikTok has fewer posts overall, but its rapid growth and high engagement rates position it as a potential hotspot for spreading harmful narratives, particularly in the lead up to the 16th General Election by 2028. While X (formerly Twitter) has an overall smaller user base, the significant volume of Level 2 and Level 3 content signals a troubling normalisation of hate speech often influenced by real-time news and trending discussions.



**Irrelevant comments were removed*

Figure 3: Snapshot of Level 3 and Level 4 severity comments on Facebook and X (formerly Twitter), August 2024

5.3 Key Actors by Severity Level:

Table 5: Key Actors by Severity Level

Actors	Total Messages	Level 1	Level 2	Level 3	Level 4
Media	2,038	1,914	112	18	0
Government Enforcement	663	561	96	5	0
KOLs	483	318	157	14	0
Potential CIBs	12	11	1	0	0
Others	27,907	21,181	3,855	485	26
TOTAL	**31,103**	**23,685**	**3,646**	**522**	**26**

Digital spaces, spanning new and legacy media, frequently became focal points for public frustration and xenophobic sentiment. Posts addressing enforcement actions, such as raids or deportation, and crime-related reporting often elicit high levels of engagement with users leveraging the comments section to escalate discussions into hate speech (labeled as Others in Table 3). The 27,907 UGCs from regular social media users, while generally not offensive in nature, reflect normalised anti-migrant and anti-refugee sentiment.

The majority of government posts (663 posts) analysed at this level are non-offensive and informational in nature, covering policy updates, enforcement actions, and public advisories concerning undocumented migrants (PATI).[33] A smaller subset of the posts, categorized at Level 2 (96 posts), contained language that could be perceived as hostile or discriminatory in nature, particularly where enforcement actions were framed in ways that risk reinforcing negative perceptions of undocumented migrants.

Enforcement discourse may indirectly legitimise xenophobia through their sustained emphasis on security and economic risks posed by PATI and refugees. Similarly, media outlets reinforce these dynamics through sensationalist headline stories of government operations, which frame refugees and migrants in a negative light and amplify public xenophobia.

Despite their smaller output, KOLs aligned with nationalist or anti-refugee ideologies play a disproportionate role in amplifying hate speech, significantly contributing to Level 2 and Level 3 severity hate speech. Finally, potential CIBs showed their activity primarily falling within Level 1 severity. Overall, the trends by key actors highlight a general public sentiment that is not supportive of the Rohingyas, refugees, or broader pendatang (migrant) issues.

33 Undocumented migrants, referred to as Pendatang Tanpa Izin (PATI) in Malaysia, are individuals who are alleged to have violated immigration laws by documentation or permits, or by overstaying their visas. Refugees and asylum seekers are often mistakenly conflated with PATI, leading to their inclusion in government crackdowns and public campaigns targeting undocumented migrants. This confusion, fueled by inconsistent public and policy frameworks, perpetuates misinformation and hinders locals from distinguishing between the two groups.

While the discussions online broadly lack overt hostility, they underscore the normalisation of unfavorable attitudes and the prevalence of anti-migrant, anti-refugee, and anti-minority rhetoric in public discourse. This normalisation risks paving the way for greater tolerance of increasingly severe and harmful narratives over time.

5.4 Doxing (Malinformation)

The CIJ team identified approximately 100 incidents of doxxing against individuals who are (or are perceived to be) PATI, which is the deliberate online exposure of a person's identity and personal information without their consent, with the intent to harass or harm them. Many of these incidents involved Malaysians recording and publicly harassing Rohingya individuals in various settings, such as workplaces, markets, or residential areas, and often disclosing specific addresses where Rohingyas reside. Furthermore, these posts frequently attracted significant engagement, with commenters encouraging vigilantism, including calls for neighborhood surveillance, reporting to authorities, and in some cases, direct incitement to violence. Many users demanded mass deportations, while others used these posts to justify acts of aggression and the genocide of the Rohingya community.

A particularly alarming example involved a TikTok post of a Malay man recording a group of Rohingya children in a residential area. In the video, the man revealed the location of the children while employing xenophobic and fear-mongering language, claiming that local neighborhoods were being 'colonised' by overpopulating Rohingya families and accusing the authorities of inaction. CIJ reported this post, along with ten other similar posts to TikTok. As a result, eight of these videos were promptly removed.

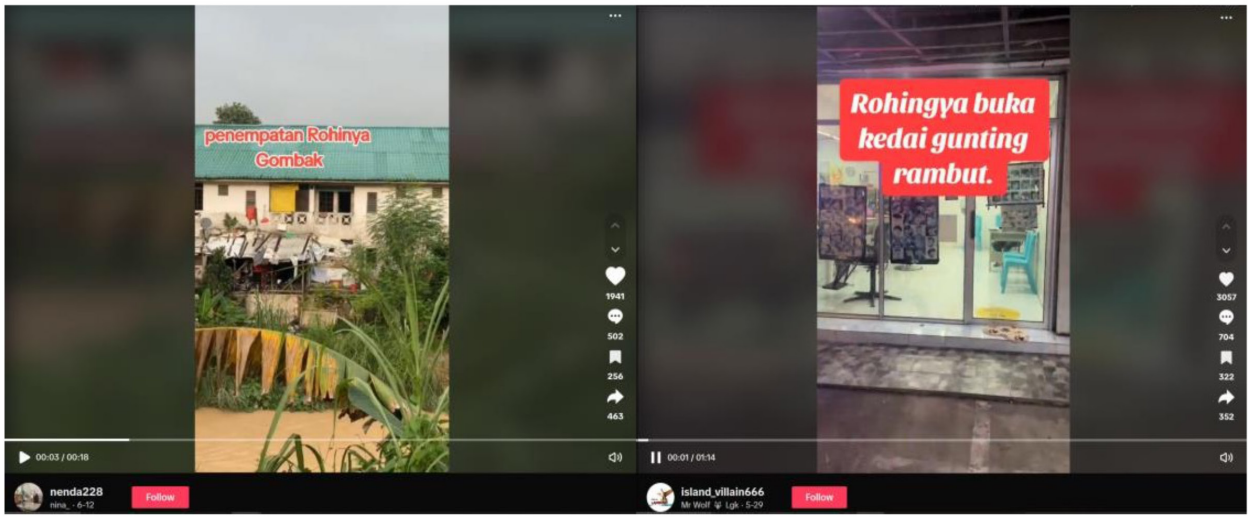


Figure 4: (Above) Snapshot of social media posts on TikTok doxxing Rohingya residents and places of commerce. TikTok, August 2024.

5.5 Coordinated Inauthentic Behaviour (CIB)

During the monitoring period, the CIJ team observed a significantly lower frequency of CIBs compared to its findings during the 15th General Election in 2022, thus this did not have any significant impact on shaping narratives about the issue monitored. This observed decline could be attributed to a variety of potential factors. Unlike high-stakes periods like GE15, where CIBs were deployed to drive highly polarised political campaigns and influence electoral outcomes, the hate speech targeting Rohingya refugees during the monitoring period appears to lack the same level of political utility. Political actors and affiliated CIB networks may have deprioritised anti-Rohingya-related narratives as it does not currently carry direct electoral consequences.

The decline in observable CIB activity could also reflect increased platform interventions by tech companies in detecting and removing inauthentic accounts and behaviour. While the lower frequency of CIBs might suggest reduced coordination, it does not necessarily indicate a diminished threat. Instead, it may be symptomatic of strategic shifts towards other politically expedient or profitable targets or even temporary dormancy until a more opportune sociopolitical period arises, which thus warrants continued vigilance in monitoring and reporting CIBs to mitigate their harmful impacts.

5.6 Disinformation-amplified Hate Speech

Disinformation operates as a powerful tool for amplifying hate speech by exploiting the interplay between key actors, societal biases, media dynamics, and algorithmic tendencies that create a feedback loop that both initiates and perpetuates hostility. To understand this process, the CIJ team examined how the cyclical relationships between the messenger, message, and messaging as interconnected drivers create a self-reinforcing ecosystem where disinformation and hate speech perpetuate one another to escalate prejudice and incite hostility.

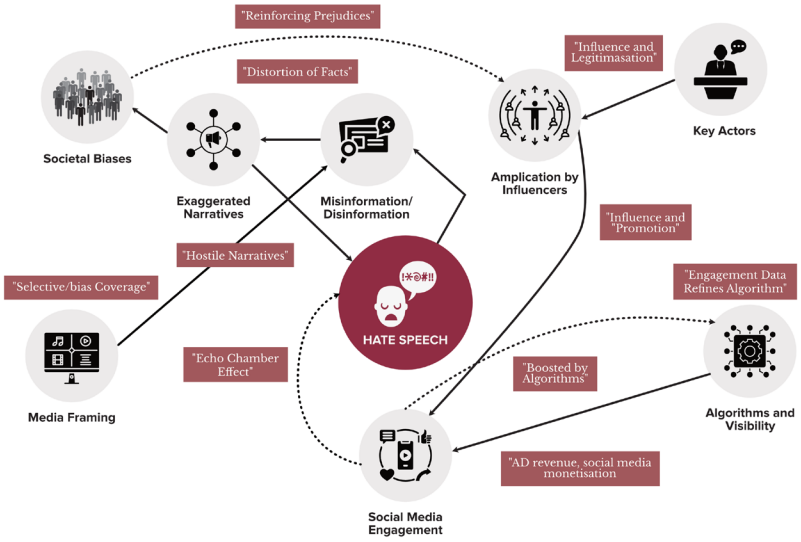


Figure 5: : Illustration of the dissemination of hate speech

1. The Messenger: Key Actors as Credible Amplifiers

CIJ observed that key actors wield considerable influence in amplifying disinformation through their platforms, acting as credible messengers whose positions lend legitimacy to fabricated narratives. Political figures, influencers, media, and anonymous networks strategically deploy disinformation tactics to introduce and validate fabricated or manipulated content into existing pools of anti-refugee and anti-Rohingya spheres.

The messenger's role becomes even more potent when their intent aligns with societal biases. Positioned as seemingly salient and plausible messages, these narratives gain traction among followers and general audiences predisposed to discrimination and have a higher deference to resonating and accepting the perspective of these messengers. Such fabricated content often spreads undetected and unchecked due to its perceived credibility when presented by these messengers. While this combination is not unique to the digital era, the internet and social media platforms have supercharged its amplification, providing rapid dissemination and unparalleled reach.

Coordinated networks of cybertroopers and bots further amplify these narratives, ensuring their rapid dissemination across platforms that create a self-reinforcing cycle where repeated exposure to disinformation can further solidify prejudices and fuel animosity.

2. The Message: Fabrication and Framing

The content of disinformation is carefully designed to exploit societal biases, using polarising and emotionally charged narratives to provoke outrage. These narratives often embed elements of truth to enhance their perceived legitimacy while creating a mix of misinformation and malinformation that is difficult for the average user to discern.

Figure 6: X (formerly Twitter), August 2024

For example, a viral social media post by user 'israhell' falsely claimed that Rohingya refugees are demanding citizenship. This narrative originally circulated in 2020 based on a letter from the Myanmar Ethnic Rohingya Human Rights Organisation Malaysia (MERHROM) to the Human Resource Minister aimed to address the challenges faced by Rohingya refugees seeking safe working conditions in Malaysia.[34] However, the original advocacy effort was deliberately misconstrued to frame Rohingya refugees as entitled and threatening the national identity. The long-lasting impact of such

34 Cilisos.my (2020, May 5). *Did Rohingya refugees actually ask for Malaysian citizenship? We check this and other claims.* https://cilisos.my/did-rohingya-refugees-actually-ask-for-malaysian-citizenship-we-check-this- and-other-claims/

disinformation is evident in its continued circulation on social media (as of the monitoring period in 2024) and its influence years after the original context has been distorted.[35]

Similarly, disinformation campaigns have falsely suggested that UNHCR-issued refugee cards provide extraordinary legal immunity, playing into existing public misconceptions to deepen distrust in refugees and humanitarian organisations. Claims that Rohingya refugees operate unlicensed businesses, despite lacking credible evidence, further exploit public fears of economic competition. These fabricated stories are designed to provoke resentment and galvanise hostility, turning public discourse into a breeding ground for hate speech.

3. The Messaging: Dissemination and Amplification

Social media platforms serve as the primary vehicle for disseminating disinformation, leveraging algorithmic biases and cross-platform functionality to spread narratives rapidly and extensively. Since algorithms designed to prioritise engagement over accuracy inherently reward disinformation that generates high levels of user interaction, key actors exploit these dynamics by repackaging and resharing disinformation narratives across platforms. This ensures sustained visibility in user feeds. Repeated exposure to disinformation creates an online environment that normalises vitriol against vulnerable groups such as minorities and refugees, particularly against the Rohingya community. This systematic recycling of false narratives not only entrenches societal biases but also fosters public hostility, which in turn reinforces the conditions for further disinformation to thrive.

This interplay between algorithmic biases, media framing, and societal prejudice ensures that disinformation is continuously circulated, recycled, and re-emphasised. Each component of this dynamic ecosystem —messenger, message, and messaging—works in tandem to sustain a self-reinforcing feedback loop where disinformation amplifies hate speech and prejudice, which entrenches a hostile environment for continued disinformation.

5.7 Implication of Normalised Hate Speech and Disinformation Online

The increasing saturation of online spaces with hate speech and disinformation targeting ‘PATI’ and refugees, particularly the Rohingya community, has profound and far-reaching consequences, both in the digital space and in real-world interactions. Repeated exposure to hate speech and disinformation desensitises individuals to discriminatory rhetoric and manufactures an online environment where xenophobia and intolerance are socially acceptable. This emboldens individuals and groups to express hostility more openly, using derogatory slurs, fearmongering, and inciting violence, acting upon their harmful biases without fear of accountability.

---

35 The statement was allegedly made by the at the time president of the Rohingya rights Malaysia organisation in 2020, however through our investigation we found that we could not trace the original statement. An article written by Free Malaysia Today clarifies that the statement of the president were blown out of porportion and he had received death threats. https://www.freemalaysiatoday.com/category/bahasa/2020/04/24/presiden-hak- asasi-etnik-rohingya-di-malaysia-nafi-tuntut-hak-kerakyatan/ The said post has been circulating since 2020 and is still being used to spread disinformation in current times.

One of the most visible manifestations of this normalisation is the widespread xenophobia found in the comments section of social media platforms such as Facebook and TikTok. Posts by media outlets and agencies announcing enforcement actions frequently attract high engagement, and subsequent user-generated comments may escalate into discriminatory or hostile speech, particularly in environments with limited moderation.

Many even go beyond expressing hostility, as some actively encourage government authorities to conduct targeted immigration raids in specific areas while others demand the expulsion of refugees.

The presence of anti-Rohingya vigilante groups and individuals, supported by online communities, further exacerbates this trend. Public figures and politicians who promote exclusionary rhetoric are often framed as 'heroes' intervening where the state has purportedly failed or turned a blind eye, thus legitimising hate-driven 'community actions' against the Rohingya.

This digital hostility has tangible, real-world consequences. CIJ monitors documented a rise in community-led harassment, forced displacement of Rohingya families, and coordinated vigilante attacks —many of which have been livestreamed on social media to garner public support. During the focus group discussions, Rohingya representatives shared with CIJ that these targeted actions are often driven by sensationalised and biased narratives that paint them as a social and economic threat. As a result, members of their community face significant barriers to basic services, particularly healthcare. Reports indicate that local residents have chased them away from government clinics and hospitals, denying them medical treatment.

There have also been instances of physical violence, with cases of Rohingya individuals being beaten by locals, as well as communities erecting banners and spreading messages demanding their expulsion from residential areas.

5.8 TikTok Engagement and the Scale of Hate

Table 6: TikTok Actors Engagement Rate

TikTok Account	Views[36]	Comments	Likes	Bookmarks	Engagement Rate (%)[37]
sophianmohdzain_official[38]	1135238	2251	33526	2280	3.35
LawanAliensKembali	564934	1932	14625	1876	3.26
SayNoToRohingyaIntruders⚔️	553100	5370	26499	2779	6.26
Utusan Rimba	4990500	9872	37535	11339	1.18
CMK_Tv	2675500	7643	99500	8764	4.33
Harimau Malaya	1777800	5364	50458	5632	3.46

The accounts identified are the top accounts identified by the automated tool Zanroo and confirmed through manual monitoring. These TikTok accounts generate massive engagement. Their top eight videos alone have gathered millions of views. In addition to views, the interaction rate (likes, comments, and bookmarks) shows these accounts are shaping opinions. Some accounts, like SayNoToRohingyaIntruders⚔️, have engagement rates exceeding 6 per cent, meaning a significant portion of viewers actively engage with their content. This matters because TikTok's algorithm rewards engagement and makes these videos available on increased user feeds. When the content spreads hate, it turns into a powerful tool for misinformation and divisive narratives.

This demonstrates that it is not just about viral content, but in actual fact, it contributes to high influence. High engagement means these accounts are not just reaching people; they are shaping narratives. That becomes a serious issue when the message is rooted in hate. Digital platforms should foster free expression, but when engagement-driven algorithms amplify harmful content, they create real-world consequences. TikTok's role in this deserves more scrutiny. It is not just about individual posts; it's about how the platform enables this content to thrive.

Another key factor in the spread of such content is cross-posting across social media platforms. A popular form of cross-posting is the TikTok-to-X pipeline. Viral TikTok videos do not stay on TikTok—if they gain traction, they often get reposted on X (formerly Twitter).

---

36 The total number of views, comments, likes, bookmarks are sampled from the top 8 post from the respective KOL account.

37 Engagement Rate=Total Views (Comments + Likes + Shares/Bookmarks) ×100 reflects how many viewers take an extra step beyond just watching—they react, comment, or save the video. A high engagement rate indicates that the content is resonating with the audience, prompting them to interact rather than just passively consuming it. Have an engagement rate of more than 5% is more than favorable for an account. https://agencyanalytics.com/kpi-definitions/engagement-rate

38 A caveat to Uztaz Sophian account being that he is a notable KOL that has had multiple accounts that were taken down during the monitoring period. As such, it would appear that he has just started a new account each time showcasing the influence from his name in spreading anti-refugee propaganda

This is especially true for videos promoting anti-refugee or nationalist narratives. Given X's more relaxed content moderation rules, such posts are less likely to be taken down. Furthermore, the platform also rewards posts with high engagement, even if it is driven by outrage or other severe emotions. Replies, quote tweets, and debates only fuel further dissemination and proliferation of hate and disinformation.

Notable accounts known for amplifying these types of videos include @Kuasaaliens and @MALAYSIAVIRALLL. These accounts brand themselves as alternative media accounts but often post content that is nationalistic and xenophobic.

Beyond social media, these narratives spread further through private messaging applications such as WhatsApp and Telegram. While the reach of such dissemination is difficult to quantify, given the limitations of the project, it adds another layer of reach and influence, expanding the potential audience even further.

06 TARGETED ANALYSIS OF KEY ACTORS

Key actors have an influential role in shaping the public narratives around Rohingya refugees in Malaysia. By examining patterns in communication, framing, and amplification across digital spaces, and by the various actors, including politicians, media outlets, government agencies, and key opinion leaders (KOLs), this report seeks to better understand the drivers of public sentiment and how misinformation, xenophobic rhetoric, and hate speech circulate online. The insights generated aim to shed light on the broader dynamics that influence public attitudes toward Rohingya communities.

6.1 Political Members and Public Figures

Certain political members and public figures have contributed to a public sphere that tolerates and even encourages the spread of toxic hate speech against refugee communities, particularly the Rohingya. These individuals, through their public statements and policy positions, reinforce the xenophobic sentiments fueling hostility towards refugees rather than appealing to the humanitarian urgency.

Statements made during parliamentary debates and by elected representatives can have a significant impact on public discourse when they circulate beyond formal settings and into online spaces.

In recent instances, migration-related discussions raised in legislative and constituency contexts were framed around concerns about crime, public order, and community impact, and were widely shared on social media.

Once amplified online, such framing generated polarised reactions, with some users echoing exclusionary narratives and portraying refugee communities as sources of social disruption or insecurity. These dynamics were further reinforced through social media content featuring public confrontations and strong rhetoric directed at refugee populations.

In March, during a session in the Dewan Rakyat (House of Representatives), a Member of Parliament (MP) called on the government to establish clear policies and guidelines to regulate the Rohingya community.[39] The MP's appeal was framed around concerns about crime and the alleged threats posed by the presence of Rohingya refugees in residential areas. It was claimed that the Rohingya "threatens the lives of the people in this country", citing more than 200 Rohingya families who have taken over a squatter settlement in his constituency to be reasons for social ills, filth, and disruption to social harmony.

---

39 Chan, M. (2024, March 7). Clear govt policy, guidelines needed to regulate Rohingya, says MP. Free Malaysia Today | FMT. https://staging-beta.freemalaysiatoday.com/category/nation/2024/03/07/clear-govt- policy-guidelines-needed-to-regulate-rohingya-says-mp/

Another assemblyman, popular on his official social media channels for his anti-Rohingya remarks, warned Rohingya refugees in his constituency that "there is a crazy YB who will always push [against Rohingyas]". Despite repeated calls from civil society groups to the Malaysian online regulator to swiftly deal with the politician's xenophobic content and the call to violence that often accompanies it, his TikTok channel is currently still active, with the occasional recorded videos of him conducting public confrontations targeting Rohingya in his constituency.

The persistence of such content online highlights the challenges associated with moderating high-visibility political speech in digital environments, particularly where user-generated commentary includes hostile or discriminatory language. These patterns demonstrate how offline political discourse, when amplified through social media, can shape and intensify public attitudes toward refugees.

6.2 Malaysian Government Agencies

Public communications and enforcement actions often shape online attitudes. Analysis indicates that such communications, while primarily administrative, informational in nature or presented as security advisories, often can, in high-engagement online environments, be interpreted in ways that reinforce negative perceptions of undocumented migrants, refugees, and asylum seekers, including Rohingya communities. References to security or economic pressures, when amplified online, contribute to narratives that portray the refugee population as risks or burdens to society, fuelling public hostility, even when no such intent is stated.

During 2024, enforcement-related announcements concerning undocumented migration received heightened visibility, particularly amid extensively publicised operations. These posts often generated large volumes of user-generated comments, with discussions sometimes escalating beyond the scope of the original announcements and including hostile or discriminatory rhetoric.

Additionally, public advisories encouraging the reporting of undocumented migration have, in several instances, coincided with the circulation of personal information, location-based claims, and other forms of harmful online content. Such dynamics highlight the importance of contextualised communication and safeguards to prevent the misuse or misinterpretation of enforcement-related information in digital spaces.

In 2024, enforcement agencies escalated efforts to crack down on undocumented migrants, including Rohingya refugees. In the first week of January alone, approximately 200 enforcement raids were carried out in Kuala Lumpur. These raids were widely publicised, further cementing the perception of refugees as illegal foreigners and reinforcing the criminalization of their presence.

Government posts also encourage the public to report 'illegal immigrants', known as pendatang asing tanpa izin (PATI). This has contributed to the spread of malinformation, misinformation, and hate speech, as it encourages the public to provide personal information and the locations of refugees.

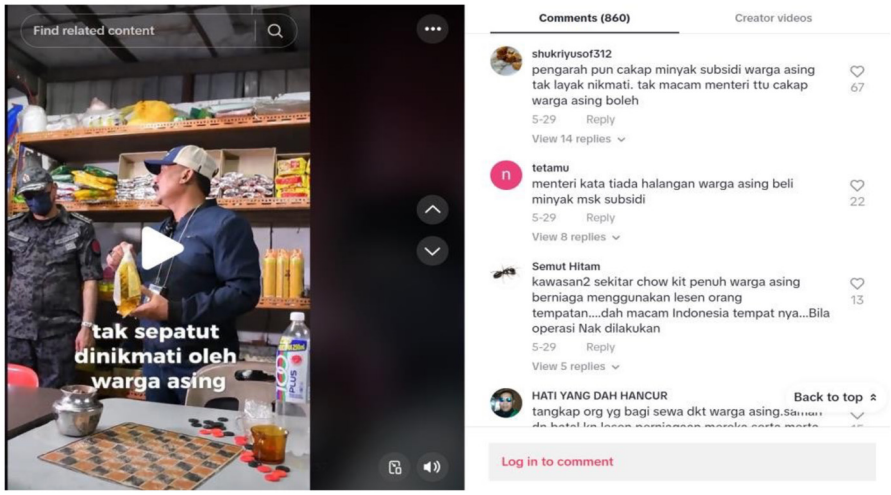


Figure 7: TikTok, August 2024

(Above) A featured TikTok post (above)[40] documented a multi-agency enforcement operation addressing alleged undocumented settlements in an urban area of Kuala Lumpur, which resulted in the detention of over one hundred foreign nationals. The video highlighted concerns regarding the distribution of subsidised goods (cooking oil) which was intended for Malaysian citizens.

The post generated high levels of online engagement, with user-generated comments largely expressing support for enforcement actions. Many responses echoed exclusionary narratives, including calls for similar operations to be carried out in other locations. Such reactions illustrate how enforcement-related content, once amplified through social media, can contribute to the circulation and normalisation of hostile narratives toward foreign nationals, including refugees and migrants. The post received 19,200 likes, 860 comments, and 1,795 shares.

6.3 The Media Ecosystem

The intersection of mainstream media, new media, and administrative messaging in Malaysia has created a media ecosystem where sensationalism and bias take precedence over factual and contextual reporting. This ecosystem not only shapes negative public sentiment but also The intersection of mainstream media, new media, and administrative messaging in Malaysia has created a media ecosystem where sensationalism and bias take precedence over factual and contextual reporting. This ecosystem not only shapes negative public sentiment but also enables the normalisation of xenophobia and hate speech targeting communities at risk like the Rohingya community.

1. Mainstream Media Encouraging Xenophobia

Mainstream media outlets frequently act as conduits for enforcement messaging, broadcasting official statements and policies on refugee issues. While these reports may appear factually neutral, their tone and framing often cast the Rohingya as outsiders who disrupt social norms. News stories linking undocumented migrants, including refugees and asylum seekers, to

40 The above TikTok post by Dato' Ruslin Jusoh, the then Director General of the Malaysian Immigration Department, showcased a joint government operation in Sentul involving the General Operations Force, National Registration Department, Malaysian Civil Defense Force, and Kuala Lumpur City Hall.

alleged crime and illegal activities perpetuate the stereotype that the Rohingya are inherently predisposed to unlawful behaviour.

Examples of media headlines such as "Colonies of Rohingya taking over different locations" by New Straits Times,[41] and Harian Metro's[42] "Conduct the whitewashing of Rohingya refugees, send them to Bhasan Char,[43] and Berita Harian's "Kulim in Kedah identified as being 'colonised' by thousands of foreigners",[44] highlights how mainstream media outlets tend to employ inflammatory language that reinforces xenophobic narratives. These narratives then spill into comment sections, where hate speech and xenophobic remarks by readers further entrench hostility towards the Rohingya.

These dynamics underscore the important role of responsible and ethical journalism in shaping public understanding of migration and refugee issues. Careful use of language, contextual reporting, and adherence to principles of accuracy, proportionality, and non-discrimination are essential to avoid reinforcing harmful stereotypes or inflaming public sentiment. Ethical reporting can help ensure that coverage informs the public without contributing to fear, hostility, or the marginalisation of communities already at risk.

2. New Media: A More Aggressive Amplifier

The new media landscape in Malaysia, dominated by platforms like Facebook, X (formerly Twitter), and TikTok, is quickly becoming central to shaping public opinion. Unlike mainstream media, which traditionally is subject to stricter editorial standards and government oversight, new media actors—including independent media outlets, bloggers, and influencers – operate with minimal regulation. This lack of accountability allows them to rapidly (re)produce and disseminate biased content that fuels anti-Rohingya sentiments.

Popular new media channels such as Malaysia Most Viral and MyNewsHub on X, and My Khabar and Malaysia Bangkit on TikTok are largely driven by clickbait, viral trends, and repackaging news stories with added sensationalism designed to provoke outrage and maximise engagement. For instance, Malaysia's Most Viral and the Vocket frequently post anti-refugee content, lamenting the so-called social burdens imposed on Malaysians by the presence of the Rohingya. MyNewsHub exaggerates or misrepresents incidents involving foreigners, sensationalising these events to draw unfounded links to illegal activities. Such narratives capitalise on emotional manipulation and lack a factual basis, presenting baseless, suggestive, and wholly irresponsible claims that go unchecked due to the absence of effective

---

41 Rahim, N.F..A., & Noh, M.F. (2024, January 21). "Colonies" of Rohingya taking over different locations nationwide. New Straits Times Online. https://www.nst.com.my/news/nation/2024/01/1003822/colonies- rohingya-taking-over-different-locations-nationwide

42 Mohd Khalid, M.K.A, (2024, January 2). Lakukan pemutihan pelarian Rohingya, hantar ke Bhasan Char. www.hmetro.com.my/mutakhir/2024/01/1047165/lakukan-pemutihan-pelarian-rohingya-hantar-ke-bhasan-char

43 Bhasan Char is a remote island in the Bay of Bengal where the Bangladesh government aims ultimately to relocate 100,000 Rohingya refugees, a 'temporary solution' that humanitarian experts is a death trap for refugees due to exposure to terrible weather conditions and lack of access to humanitarian services, including food, education, healthcare, and employment opportunities.

44 Ikhsan, M.H., (2024, January 22). *Kulim di Kedah pula dikenal pasti 'dijajah' ribuan warga asing*. https://www.bharian.com.my/berita/nasional/2024/01/1203165/kulim-di-kedah-pula-dikenal-pasti-dijajah-ribuan-warga-asing

editorial guardrails.

(Below) An X (formerly Twitter) post from MyNewsHub falsely linking a local crime spree to undocumented Rohingya refugees, despite official reports later confirming no refugee involvement. The caption states "A case of robbery and a machete attack was said to have taken place in Pasar Borong Selayang...video of victim severely slashed full of blood... watch on Channel Telegram Mynewshubviral if you wish..." The post gained significant traction with over 1.7 million views, 5,200 likes, and 2,500 shares, which also incited large numbers of xenophobic comments.

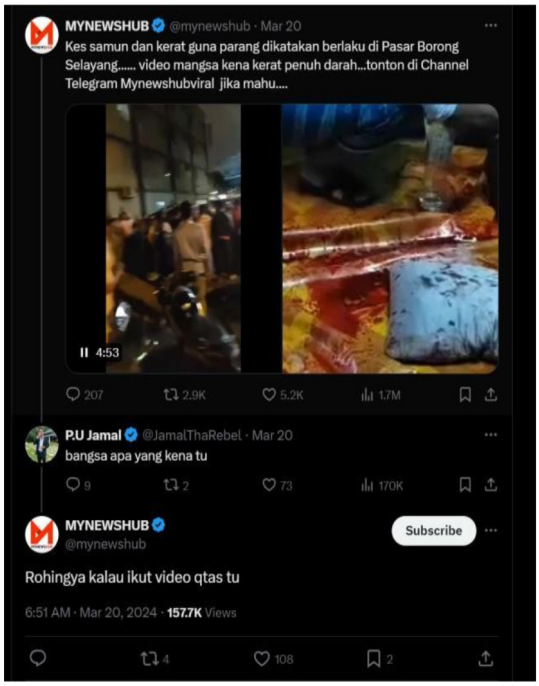


Figure 8: X (formerly Twitter), May 2024

(Below) An article from online news portal The Vocket with the headline "Malaysians regret seeing health clinics filled with Rohingyas."[45] The article cites known anti-Rohingya activist Demi Negara, who secretly filmed a Rohingya couple getting treatment at a clinic. Demi Negara's dissatisfaction was evident in his caption, "Citizens pay taxes for facilities but Rohingya are filling up health clinics using (allegedly) photocopied UNHCR letters, every year they birth more bastards. Illegal refugees stealing citizen's rights." Demi Negara's post on X, amplified by The Vocket, skyrocketed in traction reaching over 600,000 views, 3,441 reposts, and more than 4,400 likes.

Figure 9: Snapshot of an article from The Vocket, August 2024

Targeting individuals for seeking medical care is particularly harmful and deeply unethical. Accessing healthcare services does not constitute wrongdoing, and portraying refugees as abusing public facilities obscures important realities. In Malaysia, refugees generally pay significantly higher out-of-pocket fees for healthcare than citizens, who benefit from extensive government subsidies. Framing refugees as illegitimately accessing services not only misrepresents the facts, but also risks deterring vulnerable individuals from seeking essential medical treatment, with serious consequences for public health and human dignity.

45 *Rakyat Malaysia kesal lihat klinik kesihatan dipenuhi warga Rohingya*. (2024, May 10). The Vocket.com. https://thevocket.com/rakyat-malaysia-kesal-lihat-klinik-kesihatan-dipenuhi-warga-rohingya

3. Reinforcing Government Messaging

In many cases, both mainstream and new media actors frequently amplify hardline government stances on immigration. New media channels like *Malaysia Bangkit* build on these messages by adding fear-mongering narratives to official updates. When law enforcement agencies conduct raids on undocumented migrants, these outlets will frame the operations with dramatic music, sound effects, and captions to sensationalise the operations as victories against the "foreign and illegal outsiders". These posts celebrating government raids often include hashtags like #ProtectLocalJobs, #Basmi (eradicate), and #UsirRohingya (Chase away Rohingya), turning routine policy implementations into fear-mongering content.

For example, when the Malaysian government conducts raids on illegal immigrants, pages like Malaysia Bangkit and MyKhabar cover these operations extensively with inflammatory and dehumanising headlines. This type of reporting helps normalise hostility toward the Rohingya by fostering public support for discriminatory and exclusionary policies.

Such framing is deeply harmful. Sensationalising enforcement actions and presenting them as victories against dehumanised "outsiders" risks legitimising hostility and encouraging public support for discriminatory attitudes and exclusionary responses. When routine policy actions are repackaged as fear-based narratives, they blur the line between reporting and incitement, amplifying prejudice rather than informing the public. This kind of coverage not only distorts public understanding of migration issues but also contributes to an environment in which refugees and migrants are viewed as threats rather than as people entitled to dignity, safety, and due process.

Figure 10: X (formerly Twitter), August 2024

(Above) MyKhabar covered a government operation coded "Ops Bersepadu" to deport undocumented migrants in Kedah with the inflammatory headline "PATI ENEMY OF THE STATE". The post garnered 2,008 likes, 171 comments, and 161 reshares.

4. Convergence with Mainstream Media

While new media actors operate independently, their narratives frequently converge with mainstream media outlets and vice versa. Stories initially spread by unregulated new media sources on Facebook, TikTok and X (formerly Twitter) are picked up by traditional media, granting greater legitimacy and amplification to xenophobic and misleading narratives.

Conversely, new media outlets build on pre-existing narratives from mainstream media, such as by stitching or reposting content from mainstream media channels but adding a layer of sensationalism and bias to attract more views. This convergence between mainstream and new media, together with the repetitive amplification of enforcement-focused narratives, has contributed to a media environment in which sensationalism and biased framing can overshadow fact-based reporting, creating conditions in which xenophobic narratives are more likely to take hold.

For example, a neutral news story from BBC News about a minor altercation involving a local and a refugee is repackaged and reported by Malaysia Most Viral with an inflammatory headline: "Refugee violence out of control". This triggered a wave of xenophobic comments and public outrage, amplifying the issue far beyond its original context.

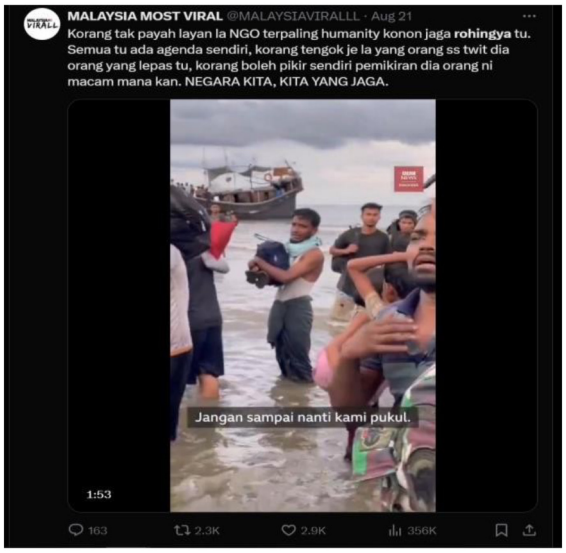


Figure 11: X (formerly Twitter), August 2024

(Above) A post from Malaysia Most Viral's X channel highlights a BBC coverage on a minor altercation between a local resident and refugees. The reshared caption reads, "No need to entertain that so-called 'most humanitarian' NGO claiming to care for the Rohingya. They all have their own agenda. Just look at the tweets people have shared about them, and you can figure out for yourselves what kind of mindset they have, right? OUR COUNTRY, WE WILL TAKE CARE.

6.4 Key Opinion Leaders (KOLs) and Influencers

The rise of key opinion leaders (KOLs) and influencers in Malaysia's new media landscape is similarly shaping the public discourse around Rohingya refugees and other foreign communities. These actors often operate under the guise of patriotism, nationalism, and social activism, with many claiming to represent the interests of the local population. This section explores some of the most influential figures and networks driving these anti-refugee narratives.

Penang Surplus Welfare Association

The Penang Surplus Welfare Association, better known online as SURPLUS Malaysia, has emerged as one of the most influential and aggressive activist groups driving anti-refugee sentiment in Malaysia. Led by Ustaz Mohd Sophian Mohd Zain, SURPLUS presents itself as a protector of Malaysian interests, claiming to defend the rights of Malaysians against the perceived economic threats posed by refugees and migrants, particularly undocumented migrants (PATI). The group asserts that foreigners, especially Rohingya refugees, areresponsible for stealing jobs, exploiting subsidies, and undermining the economy.

1. Blurring the lines of activism and vigilantism

As the leader of SURPLUS, Mohd Sophian has gained notoriety for his confrontational methods. His approach often involves "terjah" (confrontations), where he leads fellow SURPLUS activists in impromptu action in public places where they "investigate" foreigners, often targeting those working in informal sectors such as small businesses and agriculture. These confrontations, where Mohd Sophian demands to check work permits and business

licences, are frequently recorded and broadcasted across TikTok and Facebook, cultivating a reputation as vigilantes protecting local interests by framing these confrontations as defending the oppressed locals against economic exploitations by refugees and migrants.

SURPLUS actively engages local communities by crowdsourcing reports of undocumented migrants and encouraging Malaysians to participate in community surveillance. The group runs awareness campaigns on social media, urging Malaysians to act as watchdogs and tagging government agencies like the Immigration Department in their posts. Such mobilisation campaigns embolden local citizens to monitor and harass individuals based on suspicion, particularly targeting foreigners. Their tactics of combining public confrontation, sensationalised social media content, and community engagements that push a narrative of refugee and migrants as security and economic threats blurs the line between activism and vigilantism.

For example, Rohingya community representatives informed CIJ[46] that in Bagan Dalam, Penang, Rohingyas feared leaving their homes because Mohd Sopian and his team frequently barged into their shops, yelling and rudely confronting them about their businesses. They claimed that the Orang Kampung (local residents) often tipped off Mohd Sopian about Rohingya activities. Mohd Sopian would then arrive to confront them live on TikTok, seemingly aiming to go viral.

On August 15 2024, SURPLUS, escalated their efforts by visiting the UNHCR Malaysia office, purportedly on a mission to question the agency's operations. Framing the visit as an attempt to "Hold the organisation accountable", SURPLUS leveraged this confrontation to further their anti-refugee and anti-UNHCR messaging to their social media network. The visit was live-streamed on TikTok, with Mohd Sophian facing the camera and accusing UNHCR of failing to effectively regulate and resettle refugees in third countries and blaming the agency for bringing harm to Malaysian society by enabling the growing presence of refugees. The live stream was broadcast across multiple accounts belonging to prominent SURPLUS.

The TikTok broadcast format also allowed for real-time interaction, with live commenters contributing xenophobic remarks and echoing calls for the expulsion of UNHCR from Malaysia. CIJ monitors reported that approximately 600 viewers tuned in to the livestream on Mohd Sophian's channel, not including simultaneous broadcasts on other SURPLUS members' channels. The CIJ team also identified cross-platform amplification of the visitation and other SURPLUS activities, including through Utusan Rimba on Facebook and Sagato Biker on Youtube.

Content of this nature is deeply harmful. The use of confrontational tactics, dehumanising rhetoric, and public intimidation against refugee and migrant communities risk legitimising harassment and normalising vigilantism. When such actions are broadcast live and amplified for engagement, they blur the line between activism and incitement, exposing communities to fear, stigma, and potential physical harm.

---

46 The information obtained through a focus group discussion held with Rohingya refugees on 15th July 2024

The monetisation dynamics of social media further exacerbates these risks. High-engagement content driven by outrage and hostility can generate visibility, influence, and financial benefit for content creators, creating perverse incentives to escalate rhetoric and confrontation. When harmful content is repeatedly amplified and remains largely unchecked, it can contribute to an environment in which hate-based narratives are rewarded rather than discouraged.

In a multi-ethnic and multi-religious society such as Malaysia, the unchecked spread of such content poses serious risks to social cohesion. Narratives that frame particular communities as economic, security, or cultural threats can be weaponised during periods of heightened social or political tension, making them susceptible to misuse by various actors to incite division or hostility among the wider population. Preventing such outcomes requires timely moderation, accountability, and clear public messaging that rejects hate, intimidation, and vigilantism in all

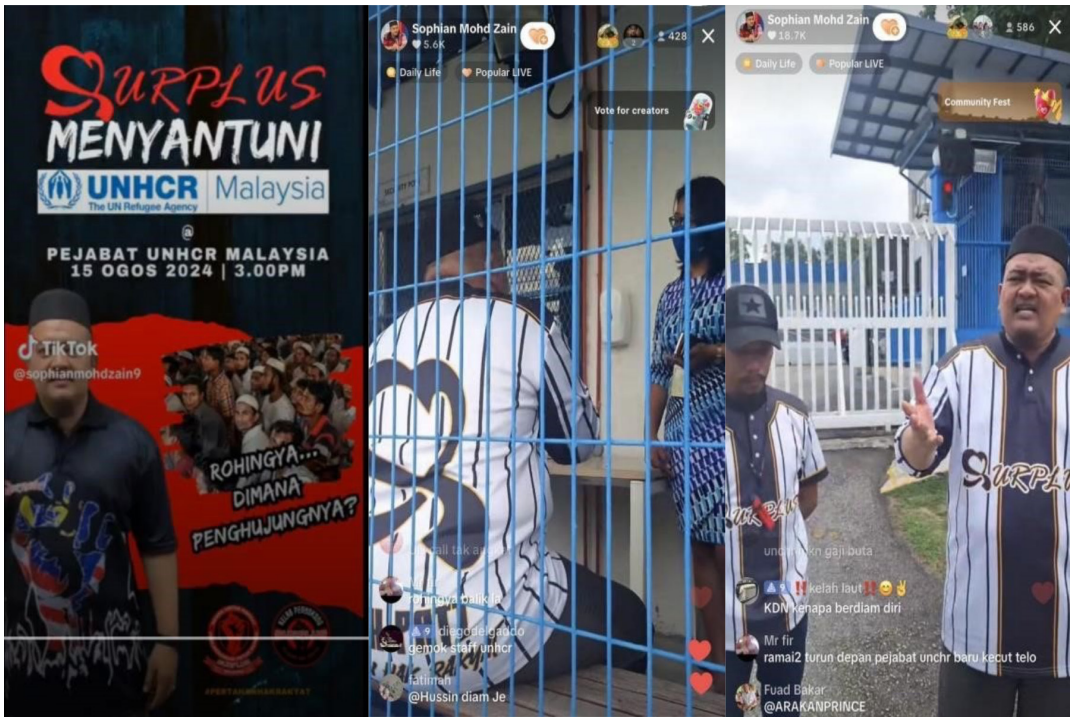


Figure 12: TikTok, August 2024

forms.

(Above left) A promotional poster shared on SURPLUS accounts on TikTok advertising their planned visit to the UNHCR Malaysia office (Above middle) A snapshot from Mohd Sophian's livestream showing him interacting with a UNHCR Malaysia staff member. During the livestream, commenters engaged in bodyshaming the staff, called for the expulsion of refugees, criticised law enforcement agencies for perceived inaction, and encouraged mass protests outside the UNHCR office, intending to intimidate the organisation. (Above right) During the same livestream, Mohd Sophian speaks to his online audience, claiming SURPLUS will hold UNHCR accountable for the lack of action in regulating refugees – particularly Rohingya - as well as expediting resettlements to third countries.

Figure 13: TikTok, September 2024

(Above) A TikTok post by Mohd Sophian highlights a 'terjah' operation in Cameron Highlands targeting foreign workers employed as cashiers in a 7-Eleven store. In his commentary, Mohd Sophian questions whether law enforcement is deliberately turning a blind eye or accepting bribes to allow such practices. This post garnered 2,287 likes, 255 comments, and 277 shares.

2. Network of Anti-refugee Influencers

SURPLUS operates as a network of influencers and activists who use social media platforms to spread anti-refugee rhetoric. Key figures like Mohd Sophian, Coachmetafizizkedry, Pejuang_Bangsa30, Razali Burhan6, Faizurrahman8520, CMK Shukrieramly, Utusan Rimba,[47] and CMKTerjah amplify each other's messages, expanding their reach and reinforcing hate speech and xenophobic sentiment. These activists present themselves as expert commentators on national policies, often criticising the government and UNHCR for their perceived failure to control the influx of refugees. Their messaging themes are often focused around these four key areas:

- **Domestic Economy:** The failure to regulate businesses employing foreign labour, which they allege contributes to unemployment and economic decline.
- **Security:** Highlighting crime allegedly committed by foreign workers and undocumented migrants, connecting these incidents to the broader narrative of danger posted by refugees and migrants
- **Culture and Identity:** Government policies that allow local Malay women to marry foreigners are viewed as a threat to Malaysian identity. Such framing reinforces patriarchal norms, undermines women's autonomy, and instrumentalises gender to advance exclusionary and xenophobic narratives.
- **Anti-UNHCR:** Calls for the expulsion of UNHCR, accusing the organisation of emboldening refugees by providing UNHCR cards that are not recognised by the Malaysian government

Through these narratives, SURPLUS activists use dehumanising language and slogans like #HapuskanPATI (eliminate undocumented migrants), #LawanPATI (fight undocumented

47 Utusan Rimba has since split from the SURPLUS network.

migrants), #SaveMalaysiaFromForeigners to reinforce their stance and mobilise public support for their content.

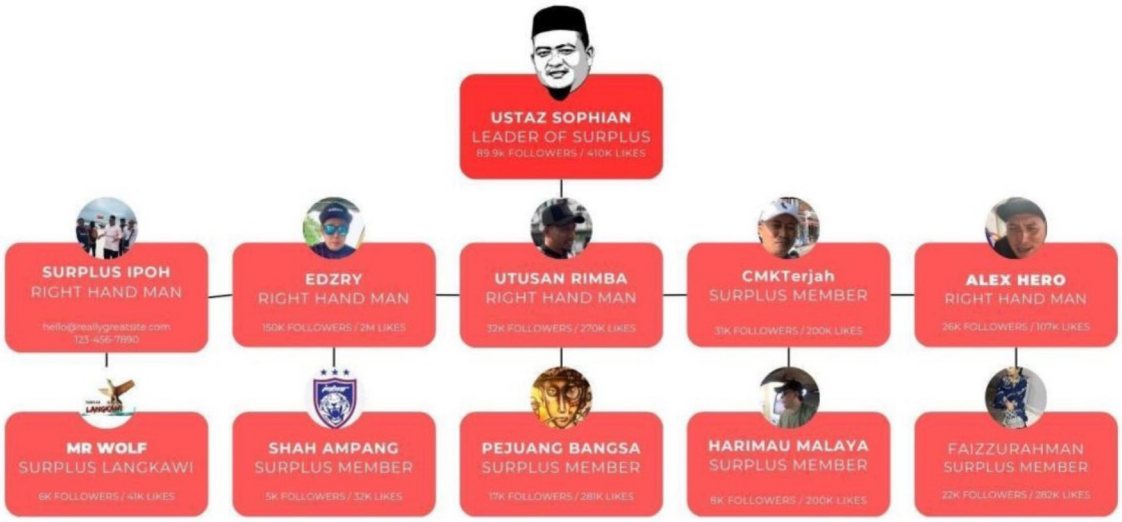


Figure 14: Family Tree Illustration of members of SURPLUS

3. Association with Government Crackdowns

A key factor in SURPLUS's rise was its ability to associate itself with the government's crackdown on undocumented migrants. As the Immigration Department and law enforcement agencies intensified their raids on undocumented workers, SURPLUS activists were quick to post videos of these raids with sensationalist headlines and captions, framing them as victories in the fight to "cleanse Malaysia of illegal elements". By piggybacking on these publicised raids, SURPLUS not only amplifies government messaging but also legitimises hate speech against these targeted groups. Furthermore, SURPLUS has skillfully used these high-profile government operations to build its own brand as a vigilante group partnering with law enforcement to protect national security.

4. Financial Incentives for SURPLUS KOLs

While the exact sources of funding support for the SURPLUS network remain unclear, several key mechanisms can be inferred based on their social media activities and public presence. Prominent figures within SURPLUS, such as Mohd Sophian, Coachmetafizikedry, and Razaliburhan67, leverage their influence and social media reach to not only spread anti-refugee rhetorics but also to market products such as tea, clothing, household items, and tech gadgets directly to their followers. This blend of vigilantism and commerce helps SURPLUS leaders to generate revenue while promoting their agenda. By consistently maintaining a strong online presence and engaging with large audiences, these influencers may also be eligible for advertising revenues on platforms like TikTok, where users can earn money based on high- traffic content.

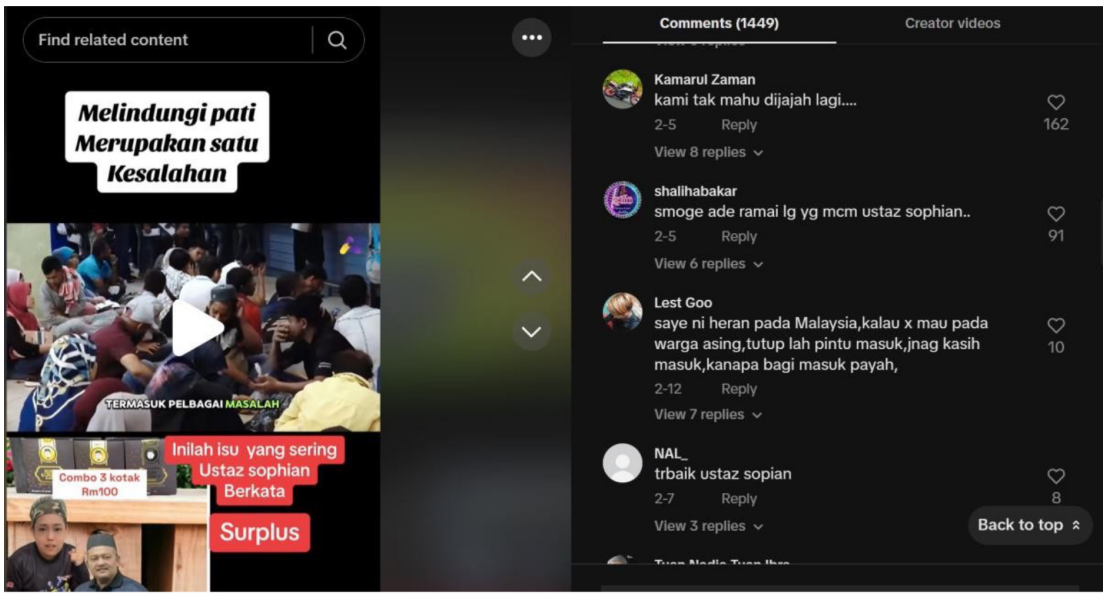


Figure 15: TikTok, July 2024

(Above) A post by Coachmetafizikedry stitching a news report from Berita 757 pushes the narrative that allowing foreigners to live in residential areas will lead to social ills and filth, leading to a threat to social cohesion. The post captioned "protecting PATI is an offence" includes an endorsement of Mohd Sophian and promotes watch merchandise for sale. The post garnered 20,800 likes, 1445 comments, and 2,814 shares.

Influencers Supporting Enforcement-Oriented Migration Narratives

Some social media influencers and advocacy-oriented online pages actively promote enforcement-focused approaches to migration and refugee issues, often framing restrictive measures as acts of patriotism or civic responsibility. Accounts like Friends of Immigration on Facebook support government crackdowns on PATI, often weaving in national security and identity themes that frame the rejection of refugees and migrants as a collective responsibility to protect the country. Two prominent figures, both awarded Youth Icon[48] titles by an enforcement agency, also use their influence, especially among young audiences, to mainstream anti-refugee rhetoric.

Analysis of such content indicates that certain narratives disproportionately target specific social groups. For example, discussions around marriages between local women and foreign nationals are often framed in ways that place blame on women for perceived threats to cultural or social cohesion. This gender-biased framing reinforces patriarchal assumptions, undermines women's autonomy, and diverts attention from broader structural and policy considerations.

48 Polis Diraja Malaysia (Royal Malaysia Police). "HARI KEBANGSAAN 2024: IKON PDRM." https://www.facebook.com/pdrmsiaofficial/posts/hari-kebangsaan-2024-ikon-pdrmbermula-hari-ini-media-sosial-rasmi-polis-diraja-m/904790605014494/.

In addition, some influencers encourage members of the public to report alleged undocumented migrants, promoting a culture of surveillance that risks normalising harassment and vigilantism. Framed as civic duty, this messaging can legitimise suspicion-based targeting of individuals and communities, particularly those already marginalised.

These influencers have also publicly endorsed proposals aimed at increasing state oversight of refugee populations, presenting such measures as necessary to improve regulation and reduce external involvement. At the same time, they downplay or dispute well-documented humanitarian concerns in places of detention, such as allegations of overcrowding, torture, and denial of food in temporary detention centres, despite credible, evidence-based reports[49] [50] from international organisations. Minimising or dismissing these concerns risks obscuring the lived realities of refugees and weakening public understanding of the human rights implications involved.

49 Human Rights Watch. (2024, January 3). Malaysia: Events of 2023. Human Rights Watch. https://www.hrw.org/world-report/2024/country-chapters/malaysia

50 Bauchner, S. (2024). "We Can't See the Sun." Human Rights Watch. https://www.hrw.org/report/2024/03/05/we-cant-see-sun/malaysias-arbitrary-detention-migrants-and-refugees

07 TRENDING DISINFORMATION NARRATIVES AND CAMPAIGNS

In the context of the Rohingya community, disinformation narratives and campaigns often intertwine with broader societal concerns, such as economic stability, national security, and cultural hegemony and identity. These narratives exploit existing social tensions and biases to present refugees, particularly the Rohingya community, as threats to Malaysia's economic resources, public safety, and ethnic harmony.

The lack of editorial oversight and fact-checking on social media allows these narratives to thrive unchecked. Content, even if factual, is often manipulated or spun to mislead audiences, further perpetuate stereotypes, entrench xenophobic ideologies, and incite hatred to reinforce systemic exclusion of the Rohingya community in Malaysia. Furthermore, these narratives are not just abstract rhetoric but have real-world consequences. This segment will show how misinformation has spurred vigilante actions, public protests, boycott calls and incitement of violence that exacerbate the challenges faced by already communities at risk groups.

7.1 Threat to Local Economy

One of the most prevalent disinformation narratives targeting the Rohingya community revolves around the belief that they are displacing Malaysians economically. The lack of a formal framework for refugees in Malaysia leaves them highly vulnerable and, over the years, has driven many to work informally in low-wage 3D (dirty, dangerous, and difficult) industries and to operate small-scale businesses. Consequently, Rohingyas are often labelled low-class migrants and illegally operating businesses, often through the alleged use of "Alibaba licences"[51] They are accused of stealing jobs from locals, despite these being jobs that locals shun away from due to social stigma and poor pay. These claims negatively frame Rohingya attempts at economic survival as acts of snatching the rights of locals, describing them as "economic refugees" with intentions of staying forever in Malaysia.

Enforcement-focused operations addressing undocumented migration have, in some instances, coincided with the intensification of exclusionary narratives in online spaces. Videos of enforcement raids frequently circulate on social media, often with sensationalist captions such as "protect local jobs", "defend Malaysia's economy", "cleanse Malaysia", and "say no to Rohingya."

These representations are often further reinforced by KOL accounts and new media channels, shaping public perceptions beyond the original context of the operations. Once amplified through high-engagement platforms, such content can contribute to simplified or misleading portrayals of refugees, asylum seekers, and migrant workers, particularly when enforcement actions are framed as symbolic victories rather than administrative measures.

---

51 A derogatory term for local business licences rented out to foreign workers and PATI to operate small businesses such as sundry shops, restaurants, and roadside traders.

In areas such as Cameron Highlands, where agriculture and tourism sectors rely heavily on low-wage labour, the visible presence of foreign workers in markets and small businesses has been cited in online discourse as evidence of economic competition. This framing risks obscuring the structural dependence of key industries on migrant labour and can reinforce narratives that depict foreign workers, especially those without regularised status, as economic threats rather than contributors to local economies.

Relatedly, social media users frequently call for boycotts of foreign-owned establishments or businesses that hire foreigners, questioning if they are qualified or "clean enough", and claiming that they are "stealing" job opportunities from locals.

In September, SURPLUS leader Mohd Sophian and his team live-streamed their visit to Medan Agro in Cameron Highlands on TikTok. The visit aimed to investigate protests by traders over significant rent hikes, with rates reportedly reaching RM5,000 per month. While the visit was ostensibly to support the trader's grievances, Mohd Sophian used his platform to push an anti-foreigner narrative, framing the presence of foreign traders as emblematic of the unfair treatment towards locals. He compared the struggles of Malaysian traders to the perceived advantages of foreigners, further intensifying xenophobic sentiment. During the livestream, Mohd Sophian stated, "This is about injustice in our own country. Foreigners are free to do business as they please without facing any consequences. When locals try to do business, they are oppressed, pressured, and outright crushed, leaving them without a source of income. Do you enjoy seeing your own people suffer in poverty?"[52]

The TikTok livestream went viral, sparking widespread online discussions and eventually drawing the attention of the Sultan of Pahang, Al-Sultan Abdullah Ri-ayatuddin Al-Mustafa Billah Shah, and the state government. The Sultan acknowledged the traders' concerns and initiated measures to resolve the issue. However, the Sultan also warned against exploiting the situation to incite tensions or deepen societal divisions.

---

52 Mimin. (2024, October 2). 35 peniaga dakwa ditindas, sewa kedai di Cameron Highlands dinaikkan dari RM500 ke RM7,500 sebulan?. The Reporter. https://thereporter.my/35-peniaga-dakwa-ditindas-sewa- kedai-di-cameron-highlands-dinaikkan-dari-rm500-ke-rm7500-sebulan/

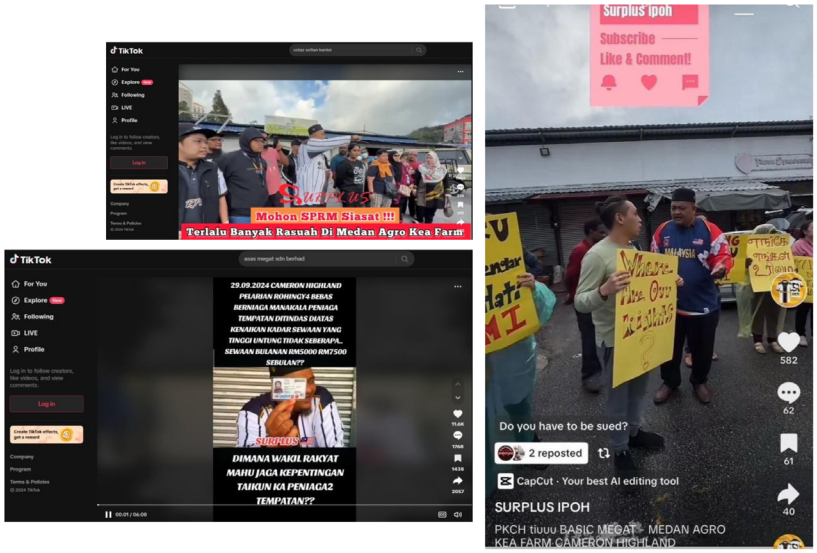


Figure 16: TikTok, October 2024

(Above) Snapshots from SURPLUS member accounts on TikTok of local traders protesting the rising cost of rental in Cameron Highlands. The video captions alleged rampant corruption in Medan Agro, and claims that Rohingyas are free to trade but local traders are slapped with rent hikes.

7.2 Threat to National Security

A persistent narrative in Malaysian discourse ties the presence of refugees, particularly the Rohingya, to a perceived rise in crime rates and problematic social behaviour, framing them as a threat to national security. This disinformation is often rooted in exaggerated or outright false claims, such as the notion that Rohingya refugees are predisposed to commit violent crimes, including the rape and murder of Malaysian women and children. While the reporting of certain incidents is factually correct—including unlawful activities committed by refugees like fights, petty crime, and confrontations with locals—the issue lies in the conflation of isolated events with sweeping stereotypes about the entire Rohingya and refugee community.

Social media has played a significant role in amplifying this narrative. Prominent social media accounts like Demi Negara and Malaysia Most Viral regularly post news articles about criminal activities involving refugees or migrants and pair them with hate-filled commentary targeting the Rohingya. These posts frequently employ fear-mongering tactics, suggesting that accepting more refugees would result in skyrocketing crime rates and a breakdown of social harmony.

Such messages are further exacerbated by new media outlets like Vocket and MyNewsHub, which consistently frame isolated cases of criminal behaviour and present them as indicative of broader societal threats, which contributes to a growing public consensus that refugees are inherently problematic and dangerous. A more constructive and humane approach to therefugee issue requires distinguishing between isolated incidents and broader community realities, and grounding public discussion in evidence rather than fear. Refugees flee conflict and persecution, and the vast majority seek safety, stability, and the opportunity to live peacefully. Framing refugee protection as a humanitarian responsibility—alongside public

order and social cohesion considerations—can help foster informed dialogue, reduce stigma, and promote responses that uphold dignity, proportionality, and shared societal values.

Figure 17: (Left) X (formerly Twitter), August 2024, (Right) X (formerly Twitter), July 2024.

(Left) A post on X by user Hafiz Adam Azlan disagreed with Comrade Confucius's criticism of Malaysians' performative outrage for the plight of Palestinians, given their xenophobic attitudes against Rohingya and the Bajau Laut people. Hafiz responded by aligning Malaysians as Palestinians under siege by Rohingya who "invaded" the country, stating "[Rohingya] rape women, kill and bury them in concrete, build illegal settlements, evade paying hospital bills. Damn it. Suddenly, defending the sovereignty of our country is Zionist behavior."

(Right) A post on X by user Demi Negara, who reshared a Harian Metro news article quoting the President of the Malaysian Consultative Council for Islamic Organisation (MAPIM). The article highlights the president's statement addressing the widespread negative perceptions of refugees among the public, emphasizing that only a small minority of refugees are involved in offenses or crimes. In response, Demi Negara retorted with an incendiary comment, "This damn MAPIM defender of Rohingya illegal intruders in Malaysia. Do they care when Rohingya rape, molest, ROB subsidies, trade, and drive without license killing Malaysians? EXPEL ALL REFUGEES FROM 50 COUNTRIES[53] IN MALAYSIA."

7.3 Zionist Colonisers

The narrative against the Rohingya is also intertwined with the misappropriation of the term 'Zionism'. Originally used to critique Israeli policies and the expansion of settlements in Palestine, the term has been co-opted on an ethno-nationalist dimension where social media users vilify and equate the Rohingya to Zionists.[54] This rhetoric portrays Rohingyas as threats to the socio-economic and cultural dominance of the Malay majority, accusing them of engaging in a form of "Zionist-like settler colonialism", exhorting baseless claims that

53 In 2022, UNHCR Malaysia reported that refugees and asylum seekers registered with UNHCR cover 50 countries fleeing war and oppression in their countries of origin.

54 Loh, B.Y.H. and Ali, S. (2024, February 6). FULCRUM. https://fulcrum.sg/rhetorical-sympathy-for-the-palestinian-struggle-in-malaysia-and-the-poignant-misuse-of-zionism

Rohingyas are demanding Malaysian citizenship, as well as allegations that their presence in Malaysia is an attempt to grow their population and ultimately usurp the Malay ethnic and cultural hegemony.

The weaponisation of 'Zionism' against the Rohingya is bolstered by the use of other dehumanising terms such as 'intruders', 'colonisers', and 'illegals'. These labels serve to frame the community as a foreign, invasive force, fostering a perception that their mere presence constitutes an existential threat. This narrative first gained significant traction in mid-2020 following a COVID-19 outbreak in the Selayang migrant market area, which housed a large population of Rohingya workers. The outbreak became a flashpoint, fueling xenophobic rhetoric and a surge of online vitriol and discrimination aimed at the Rohingya, which continues to pervade Malaysian online spaces today.

Figure 18: (Left) X (formerly Twitter), July 2024,
(Right) Facebook, August 2024,

(Left) In response to a press statement from the Zafar Ahmad Abdul Ghani, president of MERHROM calling for togetherness in combating hate speech, X user Demi Negara posted multiple xenophobic comments comparing Rohingya to Zionists, "Invade Malaysian borders and then let themselves be caught by the authorities, UNHCR comes down and give them a card to be free forever in Malaysia. The Zionists were also once refugees in Palestine." "Rohingya and UNHCR have turned Malaysia like their damn father's country, what's the difference between them and Zionists?..."

(Right) The new media channel Berita Semasa Tempatan Malaysia is spreading disinformation that Rohingya arriving by boat are not coming as refugees but as colonisers.

The toxicity of this narrative intensified significantly during the heightened Israel-Palestine conflict in October 2023. Comparisons emerged in which some Malaysians framed the Rohingya as the antithesis of Palestinians, who were lauded for their resilience and bravery in resisting oppression within their homeland, while Rohingyas were derided as cowardly and weak for fleeing persecution in Myanmar. The use of terms like 'Zionism' and comparisons between different refugee communities not only erases the extreme violence and systemic genocide faced by the Rohingya but also diminishes public empathy for their struggles while reinforcing harmful stereotypes.

In August 2024, news that over 100 injured Palestinian civilians were being transported to Malaysia for medical treatment sparked heated online debates and revealed a widespread undercurrent of anti-foreigner and anti-refugee sentiment. Public backlash against the Malaysian government quickly centred on comparisons between the treatment of Palestinian and Rohingya refugees. The controversy gained further traction when Malaysian actress and micro-influencer Izara Aishah posted a series of commentaries on X (formerly Twitter). Her posts drew sharp distinctions between the perceived courage of Palestinians and the supposed entitlement of Rohingyas, fueling an already toxic discourse and further legitimising discriminatory attitudes among her followers.

CIJ documented a surge in online activity on X (formerly Twitter) during this period, with a notable spike between 19 and 24 August. The analysis revealed that 28 per cent of all Level 3 hate speech severity (68 out of 244) during the monitoring period stemmed from this discourse, highlighting how polarising events and high-profile commentary can trigger a measurable increase in the intensity of hate speech.

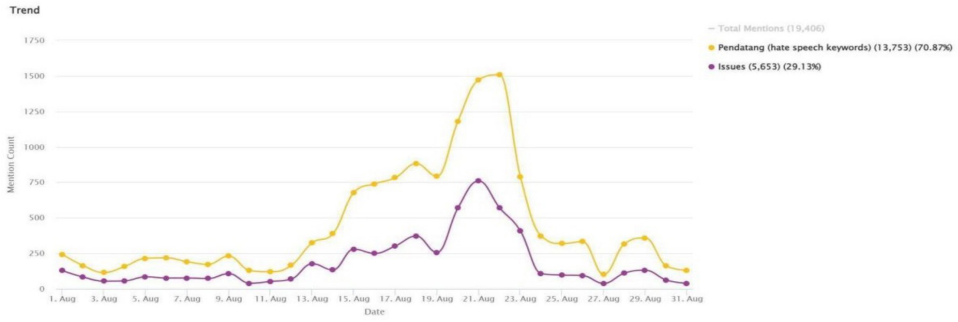


Figure 19: The Zanroo graph highlights the spike in online activity in August.

Figure 20: X (formerly Twitter), August 2024

(Top) X post by Malaysian actress Izara Aishah on August 18 insisting that the whole world is fed up with the Rohingyas and their presence, and are not qualified to be compared with the Palestinians.

(Middle) Another X post by Izara Aishah on August 18 sarcastically compared the Palestinians and Rohingya. She states, "No, no, I totally understand. One wants to give out identification cards, one wants to fight for the rights of foreign immigrants, and one is angry and wants to give citizenship. None of it is what the Palestinians who are flying here are asking for. You want to think ill of, want to tarnish the image of Palestinians who flew here for what?

(Bottom) On August 1S, Izara Aishah continued her provocative statements on X: "Here's how Malaysia and all other countries say about them. We are kind and trying our best to help. Read global news. These are facts. Malays and hypocrites like you are worse with watermelon on your name but you're offended when I defend Palestinians."

7.4 Gendered Hate Speech Against Rohingya

Rohingya women and girls endure a unique form of digital harm that heightens their vulnerability. The dissemination of false narratives, inflammatory rhetoric, and misogynistic hate speech in online discourse significantly contributes to their dehumanisation, intensifying broader anti-Rohingya narratives.

During the monitoring period, it was observed that a particular form of violent and explicit language was directed towards Rohingya women. While general hate against Rohingya is often framed by the dominant narratives that 'Rohingya people are not welcomed here', the language targeting women is distinctly sexual in nature. This reinforces harmful perceptions that Rohingya women are inferior, undeserving of dignity, or legitimate targets of abuse.

For example, the following screenshot contains highly explicit language targeted towards Rohingya women. Such a hostile environment online undermines their safety and dignity.

Figure 21: (Above) Are examples of Hate speech targeted towards Rohingya women. X (formerly Twitter), August 2024

Figure 22: TikTok, August 2024

(Above) The video, posted by the notable account @edzry, a member of the SURPLUS, makes sweeping generalisations about foreign women in Malaysia, claiming they often deceive local men into taking out loans for houses and cars to support their livelihoods. This narrative is deeply problematic as it not only reinforces harmful stereotypes but also strips foreign women of their agency and right to a dignified life. The framing of this argument is manipulative and opportunistic, as crafting a narrative with little to no factual evidence would only serve to fuel xenophobic sentiments further.

Figure 23: TikTok, August 2024

(Above) A video titled 'Rohingya kahwin melayu' posted by the account @faizurrahman8520 shows Ustaz Sophian entering a grocery store and immediately engaging in a heated argument with a woman handling the store. During the exchange, Sophian makes baseless accusations, alleging that the woman married a Rohingya man and claims that she is irresponsible to the community by selling the items at the store. This framing of marriages between Malay women and Rohingya men fuels public fear and portrays the Rohingya community as a threat to national identity. At the same time, Malay women are targeted for their personal choices, reinforcing patriarchal control over their agency and autonomy. The moral policing of Malay women implies that they are easily manipulated and should be controlled by society. This incident exemplifies how MDH not only attacks refugees but also creates another avenue to regulate and shame women in society.

08 POLICY RECOMMENDATIONS

Based on these insights, CIJ developed the following key recommendations to combat hate speech, protect communities at risk, and foster a more inclusive society:

1. UNHCR

a. Adopt policy and legal measures

i. Strengthen cooperation with the Malaysian government to develop a roadmap towards the ratification of the Refugee Convention 1951;

ii. Coordinate with government agencies, specifically the MCMC and Content Forum, to develop specific guidelines on addressing MDH targeting refugees online, particularly in line with the newly adopted Online Safety Act 2025 and the social media licensing requirements under the Communications and Multimedia Act;

b. Enhance digital and media strategy

i. Adapt and deploy an automated monitoring tool to keep track of escalating MDH against Rohingya refugees, especially during critical periods such as elections, raids, and other highly volatile situations;

ii. Collaborate with fact-checking organizations and multi-stakeholder platforms, such as JomCheck, to develop strategies and mediums to identify and debunk viral MDH about Rohingya refugees.

iii. Provide workshops to local media and influencers to ensure balanced, ethical reporting or content on refugee issues.

iv. Strengthen partnerships with tech platforms (Facebook, X, TikTok) to report hate speech and misleading content.

c. Strengthen community engagement and public awareness

i. Establish a dedicated Rapid Response Team, in collaboration with government agencies, NGOs and the Rohingya community to monitor and counter MDH with factual and timely response. This should include an alert mechanism and a crisis response mechanism to provide direct protection support for targeted Rohingya refugees;

ii. Invest in resources and time to organise town hall meetings and forums to foster understanding, dispel myths, recognise their rights as refugees and promote tolerance between Malaysians and Rohingya communities;

iii. Initiate community media projects with the Rohingya refugees as an alternate media space anchored in the lived realities of the Rohingya communities.

d. Upscale strategic communications

i. Increase proactive public messaging, in collaboration with NGOs and refugee communities, through multi-language informational materials on the status,

persecution and rights of Rohingya refugees;

ii. Engage influencers and KOLs to counter and strengthen public perception on the rights of Rohingya refugees through credible narratives and engaging mediums that can highlight Rohingya refugees' contributions through storytelling, videos, and exhibitions.

2. State

a. Strengthen Legal Frameworks

i. Set up a multistakeholder Independent Committee to address the escalation of misinformation, disinformation and hate speech. The Independent Committee should be tasked to:

- Review the root causes and drivers of misinformation, disinformation and hate speech, especially in the context of xenophobia driven by race and religion;
- Consult with civil society organisations to better understand the nature of misinformation, disinformation and hate speech and its likelihood of harm and how to proportionately respond to such acts; and
- Develop concrete actions and recommendations across different actors and platforms, and in line with international standards, specifically the Rabat Plan of Action.
- Ensure that the amended Communications and Multimedia Act and related Code of Conduct (Best Practice) for Internet Messaging Service Providers and Social Media Service Providers, as well as the recently passed Online Safety Act and other laws adequately address hate speech, misinformation, and disinformation against refugees without infringing on freedom of expression.

ii. Ensure that any measures within the legal framework is aligned with international human rights standards, such as the Universal Declaration of Human Rights (UDHR) and the International Covenant on Civil and Political Rights (ICCPR). Therefore, the principles of legality, legitimacy, necessity, and proportionality outlined in Article 19(3) of the ICCPR should be incorporated into the code of conduct to ensure alignment with the standards of freedom of expression.

b. Establish Monitoring and Reporting Mechanism

i. The Malaysian Communications and Multimedia Commission (MCMC) to establish an independent monitoring mechanism to track online and offline misinformation, disinformation and hate speech targeting refugees. This should also include providing accessible reporting tools for individuals to report incidents and seek meaningful redress safely and promptly.

c. Engage with Media and Social Media Platforms

i. The government of Malaysia should establish a Social Media Council to serve as an independent, multi-stakeholder body that addresses hate speech, misinformation, and disinformation on digital platforms. This council can create clear, transparent guidelines for content moderation, ensuring a balance between curbing illegal content and protecting freedom of expression. It would also facilitate collaboration

between social media companies, civil society, and government agencies to improve accountability and responsiveness in handling online narratives that can harm refugees. By fostering an inclusive approach, the council can help build public trust and create a safer digital environment for Rohingya refugees in Malaysia.

ii. The government, through the Malaysian Communications and Multimedia Commission (MCMC) and Jabatan Penerangan Negara, as official communications channels, to also collaborate with the media, tech platforms and fact-checking organisations to counter misinformation, disinformation and hate speech, debunk myths and amplify credible information about Rohingyas in Malaysia.

d. Promote Education and Public Awareness

i. Initiate comprehensive media and digital literacy programmes to educate and increase public awareness of the dangers of misinformation, disinformation, and hate speech, including integrating them into school curricula. Public awareness campaigns should also be initiated to counter the false and illegal narratives about Rohingya refugees while emphasising their rights and promoting equality, non-discrimination, diversity, and inclusivity.

3. Social media platforms

a. Enhance Community Standards

i. Ensure its community standards and practices are centred on human rights and corporate accountability in line with the United Nations Guiding Principles on Business and Human Rights and other international human rights standards.

ii. Platforms should include refugees and affected communities in policy development and content review teams to further enhance the approach to tackling misinformation, disinformation, and hate speech against communities at risk.

b. Localised Moderation

i. Big tech companies must invest in understanding cultural and linguistic nuances specific to Malaysia. As such, they should put adequate resources into detecting and responding to user complaints about removing misinformation, disinformation and hate-based messages on their platforms, especially during high-risk periods. Companies like Meta, TikTok, Twitter and Google can choose to divert additional resources into content moderation, including enhanced automated systems that can flag harmful language patterns.

ii. Both automated and human moderation must be available in multiple languages (specifically Bahasa Malaysia and its various dialects and nuances) and be able to contextualise and algorithmically demote hate speech to ensure it does not become virulent before removal.

c. Ensure Transparency

i. Platforms should regularly publish transparency reports detailing efforts to combat hate speech, misinformation and disinformation. Transparency reports should include data on all content moderation and removal relating to hate speech,

including those removed based on government requests. This could also support the growing demand for data from researchers.

ii. It should also include information on how platforms have revised their algorithms to reduce the amplification of misinformation, disinformation, and hate speech.

iii. Social media companies to carry out human rights impact assessments of their AI systems in place.

d. Enhance Collaboration

i. Collaborate with civil society, researchers and professional organisations on initiatives to counter hate, extremism, misinformation and disinformation.

ii. Consult researchers and CSOs in developing new tools and approaches for detecting and combating CIB.

4. Media

a. The media must guard against propagating hate speech and disinformation and refrain from giving such politics a forum or platform. It must also continue ethical and responsible journalistic practices as the standard bearer of facts and the watchdog of democracy.

b. The recent Malaysian Media Council Act 2025 must ensure that the code of conduct and related guidelines address MDH in Malaysia, including those targeting refugees.

c. Challenge and expose binary frameworks (us vs. them) that are divisive and discriminatory. Reporting should focus on exposing and calling out misinformation, disinformation and hate speech to provide solutions to stop such narratives and have more news and programmes that educate the public on digital literacy and rights within digital spaces.

d. Media can also amplify and provide counter or alternative positions to combat the proliferation of messages of intolerance or expressions that may incite violence, hostility, or discrimination towards refugees and migrants. This should also include reporting on different groups or persons who are often the targets of hate speech and allowing their members to speak and be heard in a way that promotes a better understanding of their perspectives and experiences.

4. KOL and influencers

a. KOLs and influencers play a significant role in shaping public discourse and choice. KOL must be cognisant of and draw a line between legitimate freedom of expression and misinformation, disinformation and hate speech.

b. KOLs and influencers should also take on a proactive role in condemning misinformation, disinformation and hate speech by promoting positive narratives. They can use their platforms to share stories, images, and videos that showcase the challenges and humanity of Rohingya refugees. By providing personal, relatable narratives, they can create content debunking common myths (such as misconceptions about their economic impact or alleged criminal activities) and counter dehumanizing rhetoric. This would also set a precedent for respectful and inclusive online behaviour, thus encouraging their followers to do the same.

c. KOL and influencers can also challenge algorithmic amplification, disrupting harmful trends. They can redirect attention from harmful viral content by creating counter-amplification using more inclusive and non-discriminatory trending hashtags or focused discussions.

d. KOLs and influencers can also collaborate with fact-checkers and amplify verified information to ensure their followers and audience receive accurate and reliable information. At the same time, they can promote digital literacy by sharing resources, such as guides to fact-checking or tools/links to credible sources.

09 CONCLUSIONS

This report underscores the critical need for holistic and sustainable action to address the pervasive hate speech and misinformation targeting refugees and asylum seekers, especially the Rohingya, which constitute the largest refugee community in Malaysia. It highlights how unchecked narratives can lead to severe societal consequences, including systemic discrimination, violence, and distrust in state and international institutions. The normalisation of xenophobic rhetoric in Malaysia's social media landscape has been exacerbated by insufficient legal protections and, in particular, the inconsistent application of existing immigration laws that do not differentiate among refugees, asylum seekers, trafficking victims, and undocumented migrants.

To combat these challenges, a multi-stakeholder approach is essential. Legal reforms, targeted education campaigns, and enhanced collaboration between governmental bodies, social media platforms, media institutions, and civil society organisations are vital to fostering understanding and mitigating the harm caused by unchecked hate speech. Recognising the humanity and rights of refugees is not only a moral imperative but also critical for Malaysia's social cohesion.

ANNEX 1

List of Keywords Used

Hate Speech	Pendatang	(AND) kotor	slurs, coded language
		(AND) Besar kepala	slurs, coded language
		(AND) penyakit	false claims, conspiracy
		(AND) penjenayah	propaganda
		(OR) warga asing	slurs, coded language
		(OR) pelarian	slurs, coded language
		(AND) UNHCR	false claims, conspiracy
		(AND) Bebas	false claims, conspiracy
		(OR) Refugees	slurs, coded language
		(AND) Haram	slurs, coded language
		(AND) IC	false claims, conspiracy
		(AND) Kahwin	false claims, conspiracy
		(NOT) India	
		(NOT) Cina	
		(AND) Imigresen	slurs, coded language
		(AND) Halau	slurs, coded language
		(AND) Berkerja	false claims, conspiracy
		(AND) Sindiket	false claims, conspiracy
		(AND) Tangkap	slurs, coded language
		(AND) Depot	slurs, coded language
		(AND) Selayang	false claims, conspiracy
Hate Speech	Rohingya	(AND) Kotor	slurs, coded language
		(AND) Penyakit	slurs, coded language
		(AND) Penjenayah	slurs, coded language
		(AND) Besar Kepala	slurs, coded language

	(OR) warga asing	slurs, coded language
	(OR) pelarian	slurs, coded language
	(AND) Palestine	slurs, coded language, narrative
	(OR) PATI	slurs, coded language
	(AND) UNHCR	false claims, conspiracy
	(AND) Bebas	slurs, coded language
	(AND) Refugees	slurs, coded language
	(AND) Haram	slurs, coded language
	(AND) IC	false claims, conspiracy
	(AND) Tunggang	false claims, conspiracy
	(AND) Islam	slurs, coded language
	(AND) Breed	false claims, conspiracy
	(AND) Mati	slurs, coded language
	(AND) Jajah	slurs, coded language
	(AND) Penjajah	slurs, coded language
	(AND) Tuntut	slurs, coded language
	(AND) Imigresen	slurs, coded language
	(AND) Halau	slurs, coded language
	(OR) Etnik	slurs, coded language
	(AND) Sindiket	slurs, coded language
	(AND) Tangkap	slurs, coded language
	(AND) Depot	slurs, coded language

CENTRE FOR
INDEPENDENT
JOURNALISM

Advocating Media Freedom and Access to Information