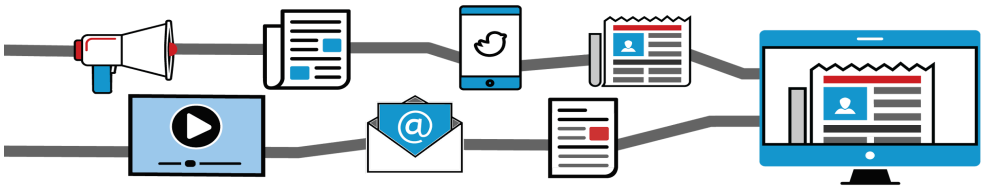


Promoting Resilience for Social Media Monitors in Malaysia Case Study: Center for Independent Journalism

The 2022 General Elections in Malaysia were held in the midst of political turmoil and democratic fallout. In the previous four years, the country saw four successive governments and four Prime Ministers, an unprecedented level of volatility for Malaysian politics. Malaysian civil society, media, and analysts anticipated intensified distortion of the information environment, particularly through hate speech and abusive narratives, in the politically contentious pre-election period. In particular, political actors have often used the tactic of weaponizing inflammatory tropes around race, religion, and royalty (also referred to as the "3Rs") for their political advantage.

In anticipation of the 15th General Elections (GE15) and subsequent local elections, the Centre for Independent Journalism, Malaysia (CIJ) developed a social media monitoring program to strengthen the resilience of Malaysian society to resist the effects of mis/disinformation and hate-based narratives in the information ecosystem.

CIJ is a non-profit organization that is working to build a democratic, just and free society where all peoples will enjoy free media and the freedom to express, seek, and impart information. They have deep expertise in media monitoring initiatives, including monitoring during the 12th and the 13th General Elections in 2008 and 2013, where they demonstrated that traditional media outlets were tilting the playing field through lopsided coverage of one political party and their narratives.

Recognizing the outsized role social media was playing in the public discourse and to safeguard the information space ahead of and during elections, CIJ designed a novel social media monitoring program in partnership with the University of Nottingham Malaysia, Universiti Sains Malaysia and Universiti Malaysia Sabah. The joint project had two objectives:

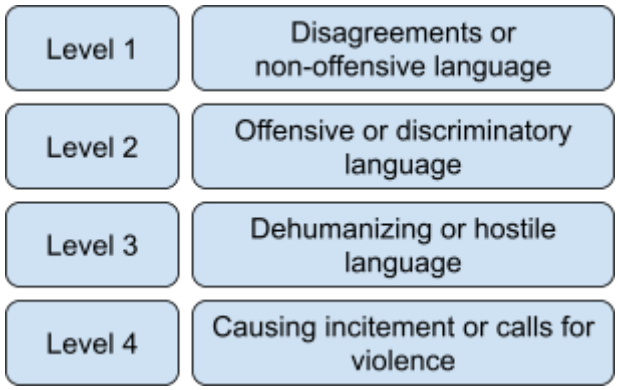
Objective 1	to expose sources and patterns of mis/disinformation, bias, and toxic rhetoric within traditional and new media
Objective 2	to develop and promote fact-based counter narratives that educate the public and hold malign information actors accountable

Social media monitoring projects in general, and specifically for an elections context, require a big investment in time, staffing and funding. This includes a sufficient project duration to reasonably recruit support staff, collect data, track narratives, and discern impacts surrounding an election. Identifying and acquiring monitoring

methodologies and tools to collect and analyze content across various platforms can be costly, and often requires testing, iterations and adjustments. CIJ proactively began planning this initiative when the possibility for GE15 to be called early became clear. Their foresight allowed the project to gain funding in time to identify the appropriate tool for their needs and hire and train monitors to conduct labeling, verification and analysis.

CIJ's Social Media Monitoring of the GE15

CIJ and Malaysian civil society's experience in previous monitoring exercises found that specific issues have increased salience during elections. Keywords and sub-themes for each issue were identified in Malaysia's three main languages; Malay, English and Mandarin. The topics they monitored were the 3Rs, gender and lesbian, gay, bisexual, transgender, intersex and queer (LGBTIQ) persons, and refugees and migrants. The monitoring also identified potential coordinated inauthentic behavior (CIB) such as bots and cybertroopers. CIJ developed a hate speech threshold scale to categorize the levels of severity as follows:



CIJ identified politicians, political parties, government agencies, media organizations, and key opinion leaders to monitor. They used a customizable tool, Zanroo, to collect content from Facebook, Twitter, YouTube and TikTok, although given API limitations, TikTok data had to be both automated and manually collected. Zanroo was able to collect quantitative data by keywords, specific pages, and character embeddings and automatically characterize content by type and actor. However, media monitors were responsible for reviewing these categorizations, along with assessing and tagging the severity level.

CIJ put together a team of monitors – mostly university students – who were then trained to review, categorize and tag the dataset according to key words and the severity level of the speech. Students also created a daily report for CIJ staff to review and identify content that potentially required rapid response from civil society.

CIJ carried out an initial pilot study from August 16 to September 30, 2022 to test the tool and review the data in preparation for use during GE15. The pilot also served to assess the readiness and reliability of the monitors, as well as the efficacy of the monitors' reporting structure. Parliament dissolved October 10 and the GE15 monitoring ran from October 20 to November 26. This monitoring period included nomination day on November 5 and election day on November 19. When the tight contest ended in a hung parliament, CIJ extended data collection for an additional week. The organization was also able to apply lessons learned and best practices from its GE15 experience for social media monitoring around the 2023 state-level elections in Malaysia.

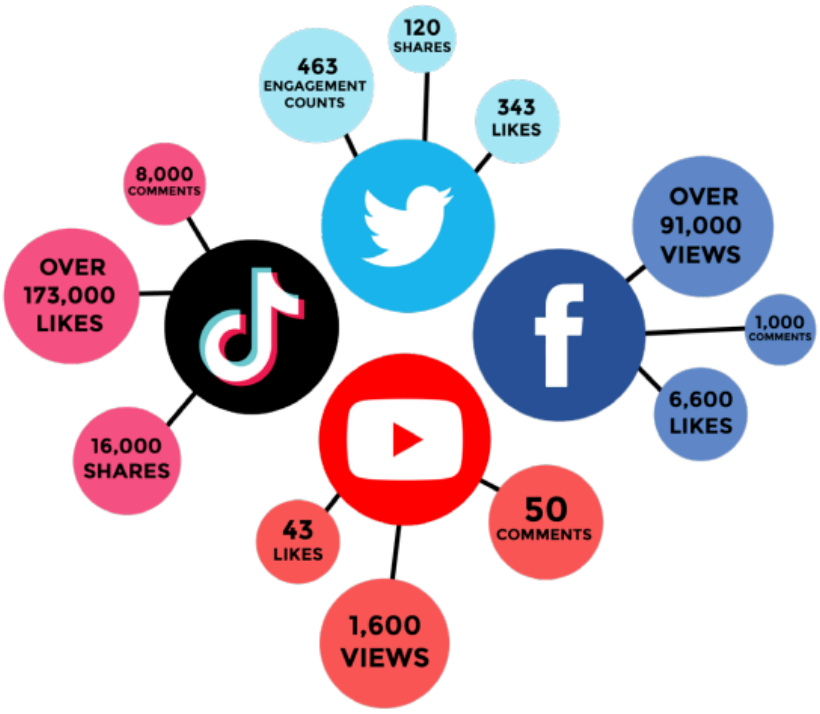
Monitoring Findings

CIJ's monitoring effort found that political campaigning online was largely conducted by influencers and celebrities, unofficial supporters, and potentially hired users, not by official party or candidate accounts. While the bulk of posts were on Facebook, followed by Twitter, TikTok emerged as a key platform for hateful videos in the period immediately following election day. The bulk of problematic posts identified were user generated content and mentions, with original posts from key political actors making up only a small portion. However, the actors also did not actively dissuade, block or take other actions to reduce the level and spread of hate speech on their pages.

Race-based narratives were most prolific in the monitoring period, however religion became the central issue in the most polarizing and divisive narratives, exploiting its intersection with a number of other social cleavages. Meanwhile comments targeting migrants and refugees were infrequent but had the highest level of severity with nearly 72% of the total level 4 posts. LGBTIQ narratives were used to undermine more liberal political factions. In addition, the monitors also noted that Malaysian traditional media generally was not proactive in fact-checking, mostly amplifying political narratives regardless of whether they were true or not.

Potential CIB was noted on both Facebook and Twitter, including the use of bots and fake accounts, although Twitter made up the bulk of CIB content. Bots and organized cybertroopers most frequently generated or boosted race or religion-based narratives. CIJ also followed how particularly viral posts, narratives and videos spread, not only within the platforms, but across them.

Chart 1



CIJ tracked the virality of a popular TikTok video that focused on race-baiting narratives.

Overall, human reviewers working on the GE15 monitoring project analyzed a staggering 99,563 unique messages, even after they were filtered for relevance and key criteria. CIJ ensured that the most impactful content was reviewed first by prioritizing posts by reach/engagement, however this still highlights the management and resource capacity challenges to reasonably process such data loads when toxic content online is nearly endless.

Table 21: Category of Actor and Platform

	PLATFORM				
CATEGORY OF ACTOR	Facebook	TikTok (automated scraping)	Twitter	YouTube	Total
Politicians	39,000	794	25,306	74	65,174
Political Parties	22,970	206	15,714	57	35,947
Media	21,454	0	7,542	125	29,121
Key Opinion Leaders	4,902	152	2,098	6	7,158
Government Agencies	3,001	62	1,533	6	4,602
Total	**91,327**	**1,214**	**52,193**	**268**	**145,002**

Comments by platform and category of actor for the GE 15 monitoring from CIJ's report

Improving Human Reviewer Resilience

CIJ set up unique staffing and workflow systems to execute their social media monitoring, labeling and analysis. Unlike some other social media monitoring projects, which may rely on a small number of full-time staff monitors, CIJ used a large pool of trained university students to review and tag content. Students worked in morning and afternoon shifts, allowing each monitor to only need to engage part-time labeling difficult content. This approach also allowed CIJ to process more data faster. However, the shift workflow also demonstrated challenges. Student attendance was unreliable, and the amount of students involved led to inconsistency in monitoring data; this was particularly true for assessing the level of severity for selected content. Additionally, given the inherent backlog with processing such large amounts of content for analysis, CIJ felt that their ability to identify and flag violating content quickly was limited, meaning problematic posts often remained online for long periods of time.

Learning from their experience, CIJ updated their workflow to improve resilience in labelers and analysts, and the sequencing of rapid response review. To try to encourage tagging consistency, CIJ placed the monitors into groups so they had support and to encourage them to talk through content they weren't sure how to address. Both mental health and response time could be improved by dividing the rapid response and data collection / analysis workflows between two different teams, creating a separate 'rapid response' team that could surface and respond to problematic content quickly rather than after it goes through analysis. In addition, CIJ worked with other civic organizations and victim service providers to respond to certain content and determine the appropriate counteraction, which could vary based on the context and vulnerabilities of populations under attack.

Social media monitors are, by design, exposed to borderline and problematic narratives that contribute to the erosion of trust in democratic institutions and societal cohesion. The work is important, but exhausting. CIJ recognized the need to protect the mental well-being of the media monitors interacting with disturbing content through specialized staff training and management, related to developing anti-trauma provisions and investing in psycho-social services and support for their monitors. CIJ developed a "Media Checklist and Support" document that provided details on how monitors can protect oneself when engaging online. They set up guidelines for monitors to jot down their thoughts and voice concerns through a clear, open channel to the CIJ staff. Initially, through this mechanism, CIJ would put people in touch with psycho-social services. In debriefs after GE15, given the impact on their mental health collectively, CIJ hired a psychosocial professional that offered a hotline service for real-time support, in-person sessions with a therapist, and access to online counseling.

Unsurprisingly, the monitors that worked more were more greatly affected. CIJ made adjustments to their staffing plan. Initially, monitors worked 3-4 hours maximum, many remotely. Monitors' resilience, labeling consistency and the volume of data they were able to review improved when CIJ requested all monitors work together in person in six hour slots, with lunch and scheduled breaks. Team leaders managed work loads, allowing more bandwidth to improve their adaptability and effectiveness by continually updating their keywords as violating users often changed their hate speech slang to evade content moderators.

Social media monitoring projects are increasingly important to analyze the information environment surrounding an election. CIJ's approach to work with university students allowed them to analyze a large amount of data, providing them with a deeper understanding of how the narratives online impacted the credibility of the GE15 election, and state-level elections that followed. CIJ recognized early on the mental and emotional burden that reviewing hateful content can create and proactively provided resources, support and care to their monitors, an area that often goes overlooked when organizations consider a social media monitoring project. CIJ's efforts offer an important case study in how election monitoring organizations can improve their social media monitoring projects and contribute positively to the methodology by prioritizing monitor resilience. Through iterative learning, review and flexibility, they were able to adjust staffing plans and work flows to improve rapid response time and foster a sense of impact in monitors. Chiefly important is that CIJ supported individual resilience by providing trainings, guidelines and counseling support services to the monitors conducting the work.