



**Monitoring online
hate speech and
incitement to
violence:
South Africa
high-risk event
30 June 2026**



Centre For Analytics
and Behavioural Change



Table of Content - 1

Summary	2
Background	2
Methodology	3
Normative framework	4
Findings and Narratives	5
Narratives	6
Dominant narratives identified	7
Reported Content and Primary Classifications	8
Platform Response and Efficacy	9
X visibility restriction	10
Priority Recommendations	11
Disclaimer	12

Summary

Between 23 June and 4 July 2026, xenophobic hate speech, mis- and disinformation, and incitement to violence were tracked across four social media platforms in South Africa. The analysis included 113 posts that were reported to the platforms. X accounted for most of the reported content, with 64 posts (56.6%), followed by TikTok with 31 (27.4%), Facebook with 15 (13.3%) and YouTube with 3 (2.7%). Platform action was recorded on only 37 items (32.7%). These findings point to a significant gap between the identification and reporting of harmful content and effective platform responses.

Background

During May 2026, South African civic movements known for their use of xenophobic hate speech and the dissemination of negative and harmful stereotypes against migrants issued an ultimatum: undocumented foreigners should leave South Africa by 30 June, and if government did not act, they would stage a national shutdown. The proposed shutdown generated significant mobilisations both online and offline. In the weeks preceding the event, calls for participation circulated extensively across social media while public statements, community organising and marches contributed to heightened public and institutional attention. This led to significant international media coverage and a statement being issued by President Rampahosa affirming the right to peaceful protest while warning against violence, intimidation, vigilantism and unlawful action, particularly in relation to undocumented immigration.

In the framework of the Social Media 4 Peace (SM4P), a project implemented by UNESCO, and funded by the European Union that seeks to strengthen the resilience of societies to online harmful content, the Centre for Analytics and Behavioural Change (CABC) tracked and reviewed content across social media platforms, and escalated content containing mis- and disinformation, hate speech and/or incitement to violence. The review took place in a period of uncertainty during the weeks preceding June 30.

The CABC treated the planned 30 June 2026 shutdown as a high-risk event (HRE) for two reasons. First, South Africa has a history of xenophobic violence against foreign nationals particularly those from other African countries. This includes a spate of violence in May 2008 when 62 individuals were killed, and tens of thousands were displaced as well as the looting of homes and foreign-owned businesses. Second was the precedent set by the July 2021 unrest when 354 individuals were killed and R50 billion was inflicted in economic damages. Like the June 30 proposed shutdown, the July 2021 unrest was preceded by explicit shutdown calls that circulated online before the major offline disruption occurred. The June 30 shutdown, however, did not develop into a comparable nationwide episode of sustained violence and looting.

Over the past six years, the Centre for Analytics and Behavioural Change (CABC) has consistently tracked online discourse and identified recurring increases in harmful and xenophobic content around successive election cycles, when issues such as immigration and access to public services become more prominent in political and public discourse. These findings provide an important evidence base for policy advocacy and targeted interventions to address online harms. As South Africa prepares for the Local Government Elections in November 2026, the report provides timely insights into emerging risks and can help inform efforts to mitigate the potential escalation of harmful narratives both online and offline.

Methodology

The process involved a first stage of tracking content in X (formerly Twitter), TikTok, YouTube and Facebook. The focus of the tracking was restricted to xenophobic hate speech, mis- and disinformation and incitement to violence on foreign nationals. This included: derogatory or dehumanizing terms directed at foreign nationals, countdown language and references to 'the 30th,' action language such as 'clean up,' 'take back,' 'community enforcement' or requests to identify targets, emojis, hashtags and regiment terminology used in a hostile or mobilization context. False official statements, manipulated media and old footage presented as current were also included. Potentially relevant content was surfaced through tool queries and in-platform searches, then reviewed by trained staff using context, target group, language, implied action and harm cues.

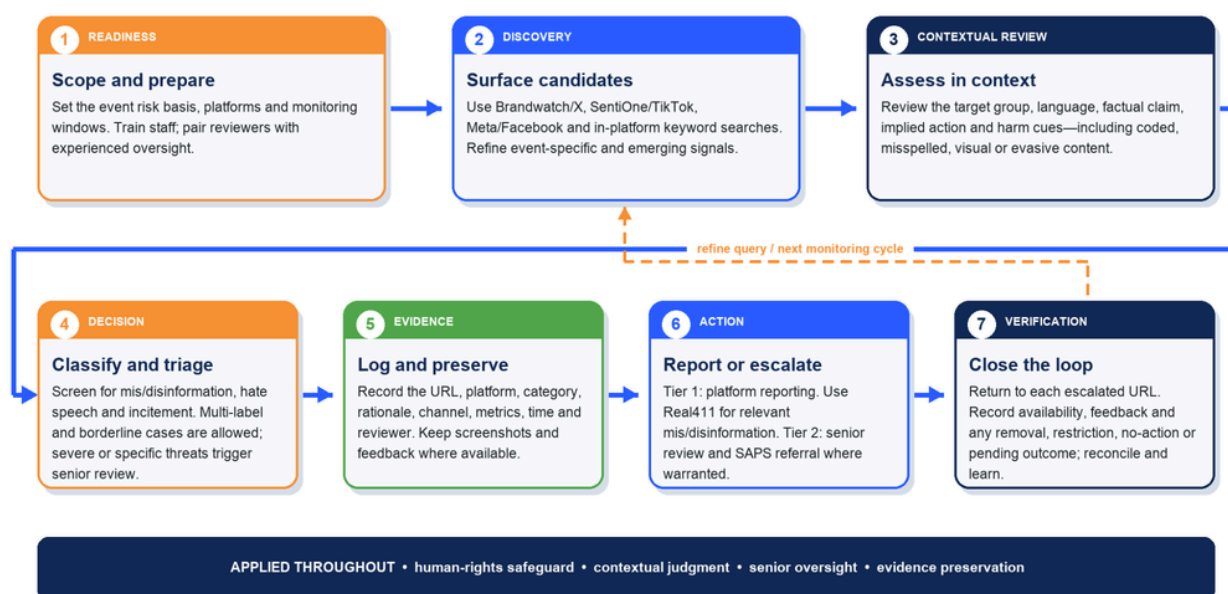
The second stage consisted of reporting the qualifying content through the relevant platform reporting channel. Content on X and Facebook was reported in-platform and content on TikTok was reported through a direct escalation channel. The assessment also documented how digital platforms responded to reported content, including platforms responses and actions taken following submissions.

The duration of the review period was from 23 June to 4 July 2026. The report was limited by the extent of platform feedback available at the reporting cut-off. An absence of recorded enforcement cannot distinguish between a pending review, a closed no-violation decision, an unavailable status update or an unrecorded outcome.

Figure 1: Process of identifying and reporting harmful content

HRE MONITORING WORKFLOW

Broad discovery is narrowed by human judgment, documented action and post-report verification.



Normative framework

Types of content identified and assessed according to the following definitions and safeguards:

Figure 2: Harmful content categories	
Content category	Definition
Misinformation / disinformation	Disinformation refers to the deliberate dissemination of false or misleading information, often intended to deceive or manipulate an individual or a group of people. Misinformation has a neutral or good underlying intent, whereas disinformation is defined as having intention to deceive.
Hate speech	UNESCO defines hate speech as communication that attacks or uses pejorative or discriminatory language against a person or group based on characteristics such as religion, ethnicity, nationality, race, gender or other identity factors. The broader UN framework specifically recognizes migration status as an aspect of a person’s identity or circumstances.
Incitement to violence	The Rabat test examines context, speaker, intent, content/form, extent and likelihood/imminence. Though guided by these criteria, the threshold used in this work was intentionally broader than what is illegal in line with international standards. We made this decision because harmful content of this category has the potential to drive offline violence despite not meeting every criterion listed in the Rabat test.
Human-rights safeguard To preserve and respect freedom of expression, criticism of immigration policy, peaceful protest, satire and news reporting were not automatically treated as harmful. The category labels listed above describe screening risk, not legal findings.	

Findings and Narratives

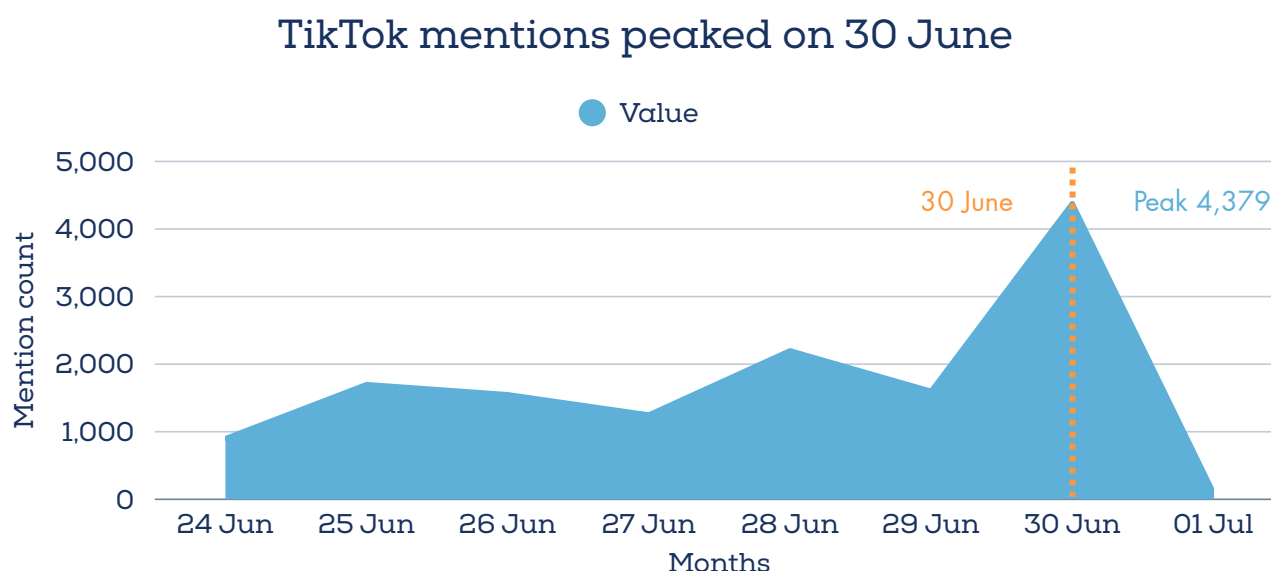
Two tool queries were used to determine the volume of mentions, reach and authors: Brandwatch for X and SentiOne for TikTok. The tracking operation in X recorded 6,722 mentions from 24 June to 1 July 2026, with 30 June as the mobilisation peak. By 19:00 that day, 1,115 mentions from 205 unique authors had generated an estimated reach of 6.79 million.

On TikTok, 13,523 matching posts were analyzed. Daily volume reached its peak on 30 June at 4,379 mentions. The term 'mentions' refer to the volume of relevant content including posts and comments while 'reach' refers to the estimated number of people/accounts that could potentially have been exposed to the total number of mentions.

Figure 3: X (formerly Twitter) metrics retrieved using Brandwatch



Figure 4: Volume of TikTok mentions over time retrieved using SentiOne



The most prevalent topic clusters on 30 June:

- 'March and March' garnered 261 mentions. This nationwide protest primarily focused on demanding the removal of foreign nationals, with a featured mention describing a "peaceful" march in Pietermaritzburg where protesters demanded foreigners leave South Africa.
- 'June 30' generated 232 mentions. South Africans expressed widespread discontent and a strong desire for change, with many indicating readiness for action on this date, as one user noted "Not even in 1994 were South Africans this united in one cause".
- 'South Africans' accumulated 174 mentions. Sentiments focused on prioritizing citizens and addressing illegal immigration, with a call for government action and a sense of unity around these concerns, echoing the sentiment that "Not even in 1994 were South Africans this united in one cause".

Narratives

Dominant narratives included collective blame, dehumanizing labels, calls for exclusion from healthcare or economic life, threats linked to the 30 June deadline, glorification of weapons and violent action.

Migration was framed as a struggle over scarce public services, employment and community safety. Foreign nationals were frequently constructed as a collective economic or security burden, while citizenship was presented as the basis for preferential access to healthcare, jobs and business opportunities. This framing supported exclusionary nationalism and legitimized direct community

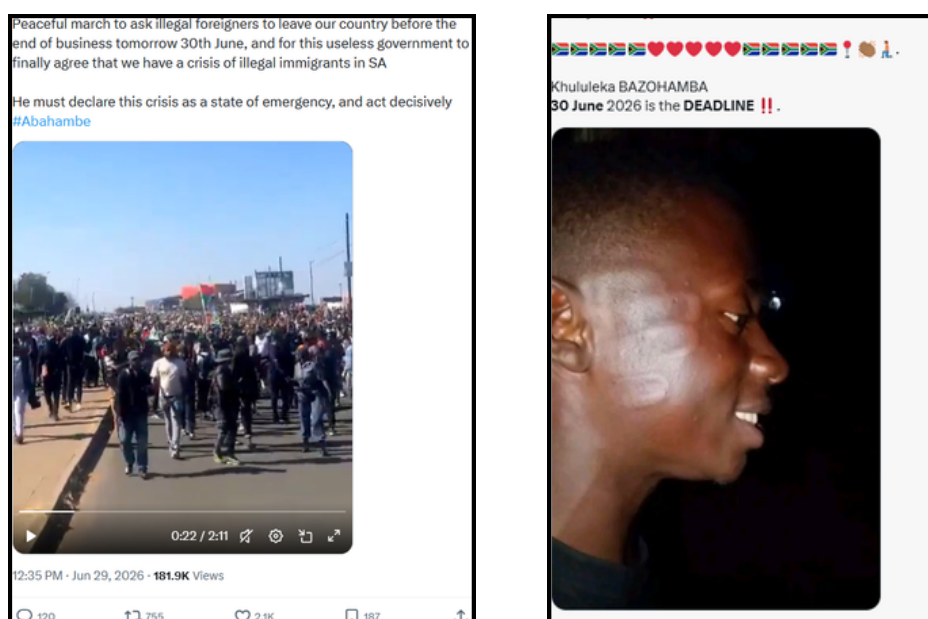
intervention where the state was perceived to have failed. It also frequently blurred the distinction between documented and undocumented migrants. Representative examples include a post describing the prioritization of South African workers as “putting our children first” and a post linking migration to public healthcare expenditure.

Further narratives presented the 30 June mobilization as an organized, legitimate citizen campaign intended to compel government action. Supporters characterized the movement as peaceful civic participation and argued that politicians and the media were deliberately misrepresenting it as xenophobic.

However, this narrative coexisted with deadline language, threats, violent imagery and calls for citizen-led enforcement, allowing coercive action to be framed as patriotism or community protection. The tension is illustrated below by a post portraying the demonstrations as a coordinated national movement, contrasted with a post encouraging violence against foreign nationals.

Content warning: The following figures reproduce xenophobic and threatening material solely for research, accountability and methodological illustration.

Figure 5: Post portraying demonstrations as a coordinated national movement (left) and post encouraging violence (right)

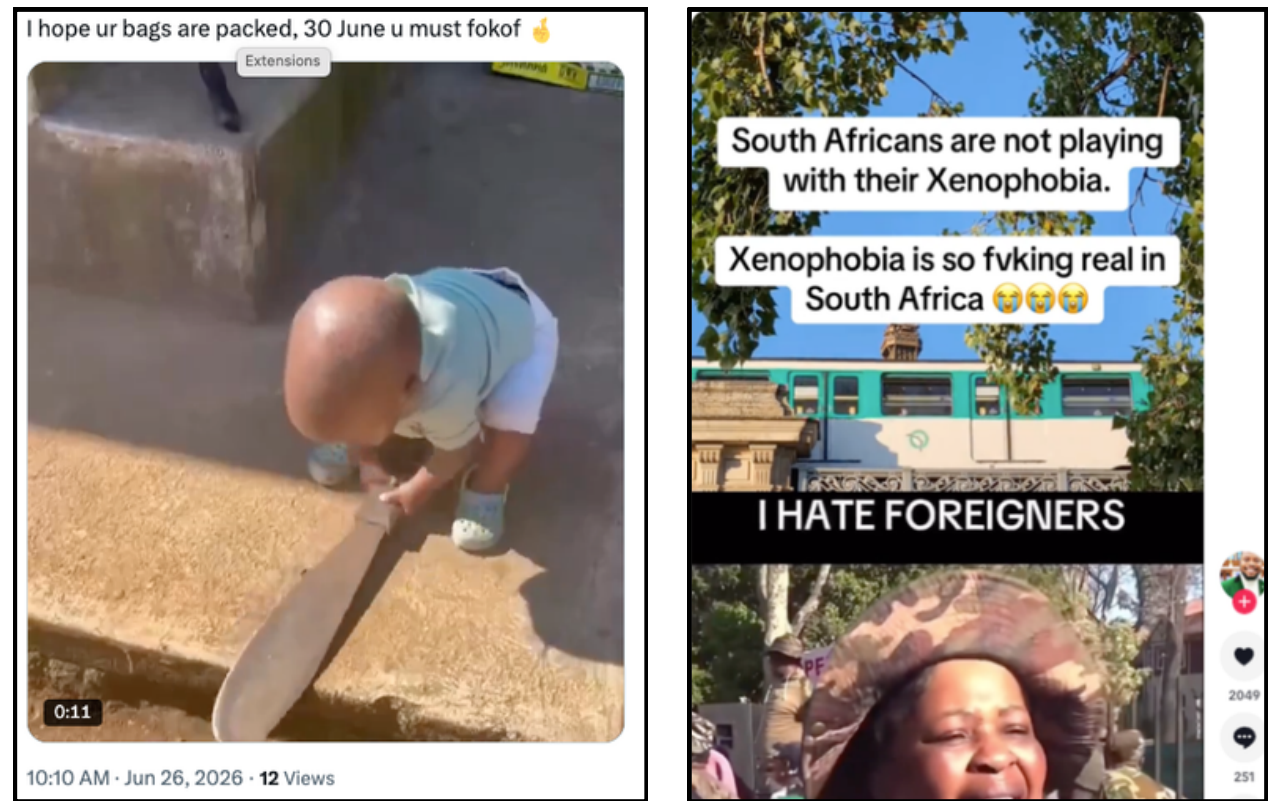


Discussions repeatedly used the phrase “illegal immigration” and framed the mobilization through calls for foreign nationals to leave South Africa, assertions of national identity and references to public action on 30 June. The South African flag commonly operated as a location or national-identity marker. Fire and other high-intensity emojis appeared within this mobilization vocabulary; their meaning remained context-dependent and was not treated as a violation without review of the surrounding post.

Figure 6: Dominant narratives identified



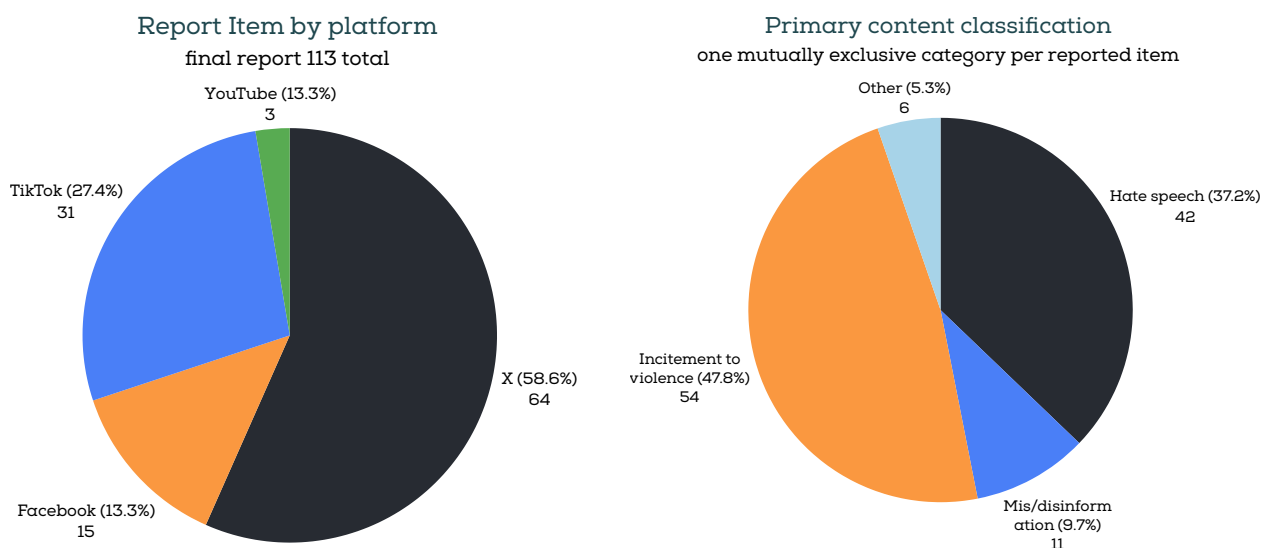
Figure 7: Examples of posts involving incitement to violence and hate speech



Reported Content and Primary Classifications

A total of 113 posts were reported for mis- and disinformation, hate speech and/or incitement to violence across four platforms. X accounted for most of the reported content, with 64 posts (56.6%), followed by TikTok with 31 (27.4%), Facebook with 15 (13.3%) and YouTube with 3 (2.7%). By content category, incitement to violence was the largest category, with 54 items (47.8%), followed by hate speech with 42 (37.2%), mis- and disinformation with 11 (9.7%), and other harmful content with 6 (5.3%). Taken together, hate speech and incitement to violence represented 96 of the 113 reported items (85.0%), highlighting the predominance of content with the potential to directly contribute to hostility, discrimination and offline harm.

Figure 8: Reported pieces of content by platform and content classification



Platform Response and Efficacy

Platform action was recorded for 37 of 113 reported items (31.8%) 36 removals or takedowns and one X visibility restriction. TikTok accounted for 30 of the 37 actions, X for seven, and no action was recorded for Facebook or YouTube. Platform reports trigger a policy review, not an automatic removal. Each platform applies its own standards, context rules and enforcement options.

Figure 9: Table showing reported items by platform with recorded action and summary

Platform	Reported items	Recorded action	Summary
X	64	7 actions	Six removals/takedowns and one visibility restriction were recorded (10.9% of X reports).
TikTok	31	30 removals	TikTok removed 30 of 31 escalated posts (97%), the strongest recorded response among the platforms.
Facebook	15	No Action	No enforcement action was recorded for the 15 reported Facebook posts at publication.
Youtube	3	No Action	No enforcement action was recorded for the three reported YouTube items at publication.

X visibility restriction:

Of the seven escalated X posts that resulted in enforcement, 6 were removed and 1 had its visibility restricted. Seen below, the visibility restricted post was hidden behind an age-restriction and a sensitive-content warning.

Figure 10: X post with reduced visibility and warning flags

Age-restricted adult content. This content might not be appropriate for people under 18 years old. To view this media, you'll need to [log in](#) to X. [Learn more](#)

Visibility limited: this Post may violate X's rules against Violent Speech. [Learn more](#)

View

Don't pay tax, bridge water, bridge electricity, hijack buildings, go to the hospital for free and so many more that makwerekwere benefit in our country. Of course not the country must burn everyone must suffer, and then we rebuild afresh working with the boers not da or anc

09:47 · 30 Jun 2026 · **199** Views

Priority Recommendations

The recommendations below follow UNESCO's Guidelines for the Governance of Digital Platforms. The guidelines recommend digital platforms to conduct human rights due diligence to address crises, situations of armed conflict, and other emergencies, including public health emergencies. This due diligence should analyze the human rights impact of the companies' operations, products, services, and advertising systems, on crisis and conflict dynamics. The findings support the following measures to strengthen platform accountability and future high-risk reporting:

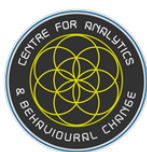
- **Establish crisis protocols and direct escalation channels:** Platforms should adopt formal protocols for high-risk events and establish dedicated escalation channels for civil-society organizations equipped to surface and escalate harmful content. These channels should provide expedited reporting, acknowledge urgent cases within 24 hours, prioritize them according to risk-based timelines, and communicate final decisions and appeal options.
- **Protect Migrant and Immigration Status:** Platforms should expressly recognize migrant and immigration status as a protected attribute. Their policies should address collective blame, dehumanization, harmful stereotypes, discriminatory exclusion and false or manipulated claims that create a foreseeable risk of harm to migrant communities.
- **Comprehensive Policy Review:** Further research should be conducted with a focus on the policies of X, TikTok and Meta and comparing the same against international standards, including UNESCO's Guidelines for the Governance of Digital Platforms, the United Nations framework on hate speech and the Rabat Plan of Action. Additional primary escalation and enforcement data over a period of months, instead of days, would also strengthen the case for policy reform. The research and findings must cover hate speech, incitement to violence, crisis-related misinformation, migrant protection, moderation outcome transparency, prevention of monetization of harmful content and emergency escalation and enforcement procedures.

Disclaimer

This Study was produced with the financial support of the European Union and UNESCO. The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO or the European Union concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries. The authors are responsible for the choice and the presentation of the facts contained in this publication and for the opinions expressed therein, which are not necessarily those of UNESCO or the European Union and do not commit the Organizations.



Social Media 4 Peace (SM4P) is a project implemented by UNESCO with financial support from the EU that seeks to strengthen the resilience of societies to online harmful content, including disinformation and hate speech, while protecting freedom of expression and access to information.



Centre For Analytics
and Behavioural Change

