



SOCIAL MEDIA MONITORING GUIDELINES FOR HATE SPEECH AND RADICAL IDEOLOGIES



FOREWORD

Since its inception, the National Cohesion and Integration Commission (NCIC) has remained steadfast in its mandate to promote equality, harmony, and peaceful coexistence among all Kenyans. However, we note that our digital landscape has transformed at an unprecedented pace, with approximately 27.4 million Kenyans now active internet users.

I have watched with concern as the digital space is increasingly weaponized to incite ethnic contempt and polarize our nation. Our research has documented over 85,000 instances of hateful content targeting ethnic, religious, and political identities in a single year.

It is in response to these evolving challenges that I am proud to present the Social Media Monitoring Guidelines for Hate Speech and Radical Ideologies. This framework is designed to safeguard constitutional freedoms while mitigating the misuse of the digital space.

We are committed to legality, transparency, and proportionality. Our goal is not to censor, but to prevent digital incitement from escalating into real-world violence. I invite all stakeholders to use these guidelines as a roadmap for responsible conduct, ensuring social media remains a space for discourse rather than division.



Dr. Daniel Mutegi Giti, PhD
Commission Secretary/CEO,
National Cohesion & Integration Commission

CORE STATEMENTS

OUR VISION

A just and equitable society living in peace, unity, and dignity.

OUR MISSION

To promote and facilitate the elimination of all forms of ethnic discrimination and proactively promote tolerance, understanding, acceptance of diversity, peaceful coexistence, and unity.

CORE VALUES:

1. **Professionalism:**

A commitment to serving clients with the utmost professionalism.

2. **Affirmative Action for the Marginalized and Minorities:**

An endeavor to undertake affirmative action for these specific groups.

3. **Respect for Diversity and Inclusivity:**

Ensuring inclusion and respecting diversity in all various engagements.

4. **Integrity:**

Providing services with the utmost integrity.

Table of Contents

Section A

I. Preamble.....	iv
II. Abbreviations and Acronyms.....	vi
III. Definitions.....	vii
1.0 Introduction	x
2.0 Objectives of the guidelines.....	xiv
3.0 Core Principles.....	xv
4.0 Institutional and Legal Framework	xvi
A. International Legal Frameworks:.....	xvi
B. National Legal Frameworks:	xvii
5.0 Collaboration with Social Media Platforms.....	xviii

Section B

6.0 Monitoring Scope and Methodology.....	19
7.0 Elements for Analyzing Hate speech and radical ideologies.....	20
8.0 Monitoring Protocols	22
9.0 Monitoring Procedures	23
9.1 Preparation	23
9.2 Monitoring Steps.....	23
10.0 Response Framework.....	25
10.1 Reporting and Oversight	25
10.2 Capacity Building	25
11.0 Ethical Considerations	26
Counter-Messaging Strategy	26
12.0 Safeguarding Measures	28

I. **Preamble**

Cognizant of the transformative power of social media in shaping public discourse, and acknowledging the right to freedom of expression as enshrined in Article 33 of the Constitution of Kenya, 2010;

Recognizing the escalating threat that hate speech, disinformation, and violent extremist narratives on digital platforms pose to national cohesion, public security, and democratic integrity

Recalling past instances in Kenya and globally where unchecked digital incitement contributed to violence, polarization, and social unrest;

Appreciating the role of digital technologies in promoting civic engagement, while emphasizing the need to mitigate their misuse;

Bearing in mind the responsibility of the State to prevent incitement to violence, protect marginalized groups, and preserve national unity;

Taking into account emerging trends on online radicalization, algorithmic amplification of harmful content, and the vulnerability of youth to extremist recruitment;

Affirming the importance of a proactive, rights-respecting, and evidence-based approach to social media monitoring;

Considering the evolving digital landscape and the growing influence of non- state actors in online discourse;

Welcoming the collaborative efforts of stakeholders including the National Cohesion and Integration Commission (NCIC), the Communications Authority of Kenya (CA), the National Counter Terrorism Centre (NCTC), the Ministry of Information and Communication, the Office of the Director of Public Prosecution (ODPP), the Office of the Data Protection Commissioner (ODPC), and other key institutions.

Determined to establish a coherent and coordinated framework that promotes accountability, upholds human dignity, and prevents the weaponization of digital platforms for hate and division.

These guidelines provide a comprehensive framework for the ethical and lawful monitoring of social media platforms to detect, prevent, and respond to hate speech and radicalization, anchored in the principles of legality, necessity, proportionality, transparency, and multi-stakeholder cooperation.

II. Abbreviations and Acronyms

CA -	Communication Authority
DCI -	Directorate of Criminal Investigations
NCIC -	National Cohesion and Integration Commission
NCPWD -	National Council for Persons with Disabilities
NCTC -	National Counter Terrorism Centre
ODPC -	Office of the Data Protection Commissioner
ODPP -	Office of the Director of Public Prosecution
PWD -	Persons Living with Disabilities

III. Definitions

Coded speech - Language which uses seemingly benign terms to refer to targets of hate speech without mentioning them explicitly, in order to avoid detection and censorship.

Deep fake – a digital representation of a person, including appearance and voice that has been artificially generated or manipulated to create a realistic but false depiction of them doing or saying something.

Digital mainstream media refers to established news organizations that share content online through websites, social media, and apps.

Digital platforms - include social media platforms, messaging applications, video streaming, podcasts, websites, blogs, and vlogs.

Emerging Technologies - include technological advancements such as AI, machine learning, internet of things, block chain, big data analytics, among others.

Ethnic contempt - Words intended to incite feelings of contempt, hatred, hostility, violence or discrimination against any person, group or community on the basis of ethnicity or race.

Extremism - the holding of extreme political or religious views; fanaticism

Hate speech –

- a) use of threatening, abusive or insulting words or behavior, or
- b) displays any written material;
- c) publishes or distributes written material;
- d) presents or directs the performance the public performance of a play;
- e) distributes, shows or plays, a recording of visual images or provides, produces or directs a programme, which is threatening, abusive or insulting or involves the use of threatening, abusive or insulting words or behavior commits an offence, if such person intends thereby to stir up ethnic hatred, or having regard to all the circumstances, ethnic hatred is likely to be stirred up.

High risk - is content that poses an immediate threat to public safety or cohesion by directly inciting violence, glorifying extremist acts, or seeking to recruit or

radicalize others. It often includes dehumanizing or genocidal language, explicit calls to harm individuals or groups, promotion of violent extremist ideologies, coordinated campaigns, and is especially dangerous during times of political or ethnic tension.

Indoctrination is the process of inculcating (teaching by repeated instruction) a person or people into an ideology, often avoiding critical analysis.

Keyword filtering - Using specific words and phrases to filter content automatically, often to identify or block certain types of content.

Low risk - refers to content that may be offensive, prejudiced, or insensitive but does not incite violence, hatred, or discrimination. It often includes stereotypes, slurs, or jokes that lack a direct call to action, as well as generalized expressions of frustration or bias. This type of content typically has low visibility—receiving few likes, shares, or followers—and shows no indication of coordination or targeted messaging.

Metadata is information that describes other data—such as the time, location, device and engagement details of a social media post. It helps track patterns, identify sources and detect coordinated activity.

Moderate risk - refers to content that features explicit prejudice, incitement, or discriminatory language with the potential to escalate into hate or violence, especially in volatile or sensitive contexts. It may call for exclusion, use coded language with radical undertones, target specific individuals or groups, and often appears in spaces with a history of hate or radical content. High engagement levels can also indicate greater potential for harm.

Radical ideologies - belief systems that advocate for fundamental, often extreme, changes to existing political, social, religious, or economic structures—typically rejecting mainstream values, systems, or authorities. They may promote non-democratic, intolerant, or violent means to achieve their goals.

Radicalization - the process of adopting or promoting an extreme belief system

for the purpose of facilitating ideologically based violence to advance political, religious or social change.

Uniform Resource Locator (URL) - is the web address or link that directs to a specific piece of content online, such as a social media post, video, profile, or website.

Violence – the use or the threat of physical force, intended to hurt, damage, or kill someone or destroy something that is motivated, justified, or enabled by radical ideologies.

Violent extremism - beliefs and actions of people who support or use violence to achieve ideological, religious or political goals.

Vulnerable groups – individuals or communities more likely to be targeted or harmed more by hate speech or radical content due to factors including, but not limited to; ethnicity, gender, disability, age, sex, religion and social status. They require special protection and sensitivity in monitoring and response efforts.

1.0 Introduction

In Kenya, ethnic tensions have been an ongoing issue between and among various communities since independence, and the post-election violence in 2007-2008 was the zenith of deep-rooted historical injustices that had caused ethnic intolerance and tension.

The National Cohesion and Integration Commission (NCIC) was established in 2008 as one of the Agenda Four Commissions created by the National Peace Accord out of the realization that long-lasting peace, sustainable development and harmonious coexistence among Kenyans required deliberate normative, institutional and attitudinal processes of constructing nationhood, national cohesion, and integration.

The NCIC is mandated to “facilitate and promote equality of opportunity, good relations, harmony, and peaceful coexistence between persons of the different ethnic and racial communities of Kenya, and to advise the Government on all aspects thereof.” In line with the Constitution of Kenya 2010 which provides for freedom of expression that may be limited in instances of hate speech, propaganda for war, and incitement on the basis of ethnicity, the NCI Act 2008 provides for the offences of hate speech and ethnic contempt and their attendant penalties.

Ethnic contempt and inflammatory rhetoric in Kenya are prevalent in traditional spaces such as political rallies, public forums, and broadcast media. However, in the last decade, the Kenyan digital ecosystem has grown immensely, transforming how citizens engage, express themselves, and participate in national conversations. The first sign of this is in the trends in mobile (SIM) subscriptions and penetration rate.¹

¹ UNESCO. 2021. Ethnic Hate Speech in Kenyan Online Spaces. Paris: UNESCO. <https://en.unesco.org>.

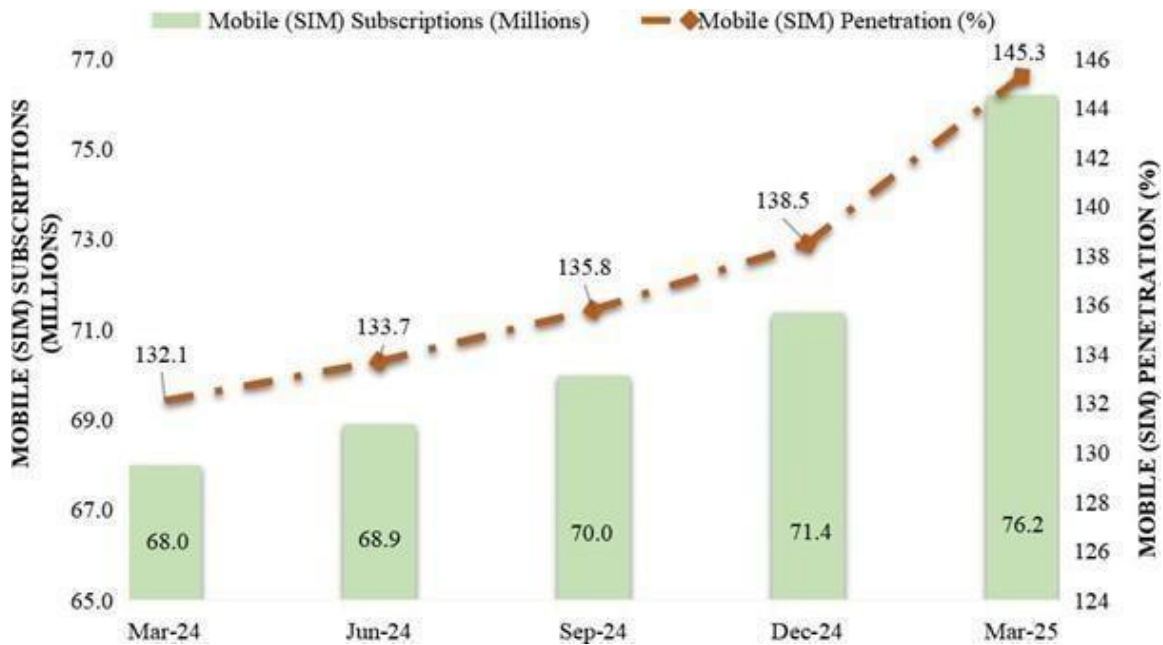


Figure 1. Source: CA, Operators' Returns, Provisional Data for Telkom Kenya Limited

As of January 2025, approximately 27.4 million Kenyans, representing 48% of the population, were active internet users, with an estimated 10 million Facebook users, 9.3 million users on YouTube, 2.5 million users on Instagram, and growing numbers of users on TikTok, X (formerly Twitter), Snapchat, and LinkedIn.²

² Standard Media. "How Online Hate Speech Threatens Peace and Kenya's Social Fabric." Standard Media, May 8, 2018. <https://www.standardmedia.co.ke/article/2001515618/how-online-hate-speech-threatens-peace-and-kenyas-social-fabric>

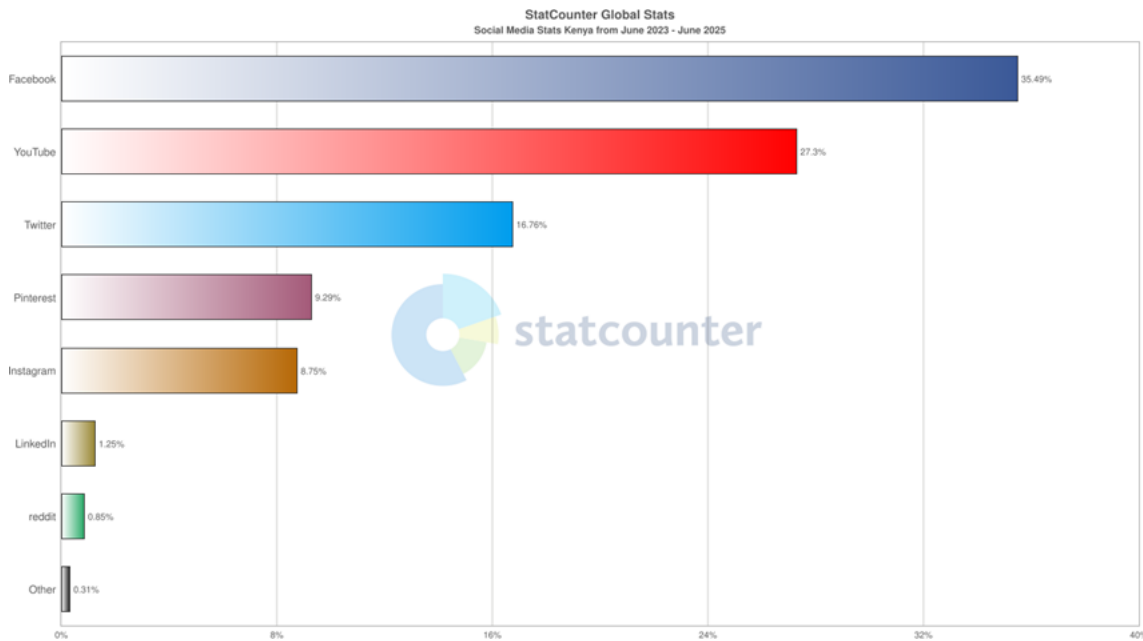


Figure 2. Social Media market share statistics for Kenya. *Source: Stat Counter*

Social media platforms are now particularly influential in the country and have become primary vehicles for communication and information sharing. Notably, Kenya ranks among the top five countries globally in terms of daily time spent on social media, with over 54% of Kenyan TikTok users engaging regularly.³

This digital penetration has thus provided a powerful platform for the spread of hate speech, incitement, and radical ideologies. For context, between May 2019 and May 2020, researchers documented over 85,000 instances of hateful content targeting ethnicity, religion, and political affiliation across Kenyan digital platforms.⁴

³ DataReportal. 2025. Digital 2025 - Kenya. <https://datareportal.com/reports/digital-2025-kenya>

⁴ Institute for Strategic Dialogue (ISD). 25th November 2021. "Online Extremism Index: Kenya": https://www.isdglobal.org/digital_dispatches/online-extremism-index-series-looking-ahead-to-the-2022-elections-in-kenya



During the lead-up to the 2022 general elections, for instance, the NCIC reported a 20% increase in online hate speech cases, with TikTok emerging as a significant vector, accounting for 20 of the 48 weekly cases, representing a dramatic 400% spike compared to previous weeks.⁵ Furthermore, in the first half of 2024, the NCIC flagged 95 hate speech cases across social media platforms, with X (formerly Twitter) accounting for 72 cases, followed by Facebook and TikTok.⁶ These figures highlight the growing persistence and complexity of online hate, driven by political incitement, ethnic polarization, misinformation, and radical narratives.

Kenya's 2007–2008 post-election violence demonstrated how inflammatory speech can escalate into real-world conflict, resulting in deaths, displacement, and deep

⁵ Wired. 2022. "TikTok's Role in Kenyan Election Tensions." <https://www.wired.com/story/kenya-facebook-elections-hate-speech-ban/>

⁶ Kenyans.co.ke. "NCIC Flags X Social Media Platform for High Number of Hate Speech Cases." Kenyans.co.ke, January 18, 2023. <https://www.kenyans.co.ke/news/101765-ncic-flags-x-social-media-platform-high-number-hate-speech-cases>

societal fractures. With the digital transformation and proliferation of social media platforms, the avenues for the spread of hate speech and radical ideologies have been broadened. Digital spaces have now emerged as the key arenas for hate and radicalization hence the necessity for the NCIC to enhance its ability to identify and respond to these risks while safeguarding constitutional freedoms. It is against this backdrop that these guidelines have been developed.⁷

2.0 Objectives of the guidelines

The overall objective of these guidelines is to provide clear, actionable procedures and legal boundaries for monitoring and responding to hate speech and radical ideologies on social media, thereby preventing incitement, promoting national cohesion, and safeguarding public order in Kenya.

They include;

1. **Establish a standardized framework** for monitoring, detecting, and addressing hate speech and radical ideologies across social media platforms in Kenya.
2. **Strengthen NCIC's capacity** for evidence-based digital monitoring, early warning, and inter-agency coordination to prevent escalation of online incitement into offline violence.
3. **Promote ethical monitoring practices** that safeguard freedom of expression, data privacy, and national values.
4. **Enhance public awareness and accountability** by guiding responsible online conduct among political actors, media, influencers, and the general public.

⁷ Statista. 2025. "Internet Usage in Kenya – Statistics & Facts." <https://www.statista.com/topics/11395/internet-usage-in-kenya/>

5. **Capacity building of actors and stakeholders** on monitoring practices against hate speech and radical ideologies.

3.0 Core Principles

- i. **Legality and Due Process:** Monitoring of social media platforms must be grounded in national and international legal frameworks, ensuring that all actions taken are lawful, authorized, and in accordance with established legal procedures, including obtaining proper warrants or approvals where necessary.
- ii. **Proportionality:** Enforcement measures should strike a fair balance between combating harmful content and safeguarding the constitutional right to freedom of expression. Actions taken must be necessary, targeted, and commensurate with the level of harm posed.
- iii. **Transparency and Accountability:** Monitoring processes must be subject to public scrutiny and institutional oversight. Agencies involved should maintain clear documentation, publish periodic reports, and provide avenues for redress and appeals where appropriate.
- iv. **Non-Partisanship:** All monitoring efforts must be conducted impartially and without political bias. The selection of content for review and enforcement actions must be based solely on objective criteria and not influenced by political, ethnic, or religious affiliations.
- v. **Privacy and Data Protection:** Personal data collected during the monitoring process must be handled with strict confidentiality, in compliance with the Data Protection Act and relevant regulations. Data should only be used for its intended lawful purpose and safeguarded against misuse or unauthorized disclosure.
- vi. **User Empowerment:** The public, especially vulnerable groups, should be equipped with the knowledge and tools to recognize, report,

And counter hate speech and radical content. Promoting digital literacy and responsible online behavior is key to building societal resilience against harmful narratives.

4.0 Institutional and Legal Framework

1. **Lead Agency:** The National Cohesion and Integration Commission (NCIC) will be responsible for monitoring hate speech, while the National Counter Terrorism Centre (NCTC) will be responsible for monitoring early signs of radicalization.
2. **Supporting Agencies:**
 - a. Communications Authority of Kenya,
 - b. Directorate of Criminal Investigations (DCI) Cybercrime Unit,
 - c. Office of the Director of Public Prosecutions (ODPP),
 - d. Judiciary,
 - e. Tech Companies,
 - f. Social Media Companies,
 - g. Fact Checking Companies,
 - h. ICT Authority,
 - i. Office of the Data Protection Commissioner (ODPC),
 - j. Witness Protection Agency,
 - k. Research Institutes & Think Tanks,
 - l. National Police Service,
 - m. Ministry of Interior and National Coordination,
 - n. Media Council of Kenya,
 - o. Kenya National Commission on Human Rights (KNCHR),
 - p. National Gender and Equality Commission (NGEC),
 - q. State Department for Children Services (DCS).

Legal Instruments

A. International Legal Frameworks:

- i. Universal Declaration on Human Rights (UDHR) and Optional Protocols;

- ii. International Covenant on Economic, Social, and Cultural Rights (ICESCR) and Optional Protocols;
- iii. International Convention on the Elimination of All Forms of Discrimination (ICERD);
- iv. General Data Protection Regulation (GDPR);
- v. Camden Principles on Freedom of Expression;
- vi. International Convention on Civil and Political Rights (ICCPR).

B. National Legal Frameworks:

- i. Constitution of Kenya,
- ii. Children's Act 2022,
- iii. Computer Misuse and Cybercrimes Act,
- iv. Data Protection Act,
- v. Evidence Act,
- vi. Kenya Information and Communications Act (KICA),
- vii. Media Council Act,
- viii. National Cohesion and Integration Act (NCI),
- ix. National Gender and Equality Commission Act.
- x. National Government and Administration Act,
- xi. NPS Act,
- xii. ODPP Act,
- xiii. Penal Code,
- xiv. Persons With Disability Act,
- xv. Prevention of Terrorism Act (POTA),
- xvi. UNDP Strategy and Action Plan against Hate Speech 2024,
- xvii. Victim Protection Act,
- xviii. Witness Protection Act,

5.0 Collaboration with Social Media Platforms

NCIC in collaboration with the Communications Authority of Kenya, will engage with social media platforms and tech companies to strengthen cooperation and coordination in the identification, addressing, reporting and response to hate speech and radical ideologies online. This collaboration shall focus on establishing clear liaison and reporting mechanisms, enhancing timely response to flagged content and promoting locally informed content moderation that takes into account Kenya's socio-cultural and linguistic content.

All engagements and monitoring activities will be conducted in compliance with the Constitution of Kenya 2010, the Data Protection Act 2019 and subsequent regulations, and be guided by the principles of legality, proportionality, respect for privacy and protection of freedom of expression in the interest of national cohesion and peace.

6.0 Monitoring Scope and Methodology

- I. **Monitoring the use of digital platforms:** to flag activities likely to undermine the peaceful coexistence of communities living in Kenya.
- II. **Commonly used digital platforms,** including Facebook, X (formerly Twitter), TikTok, WhatsApp (public), Snapchat, Telegram, YouTube, blogs, and websites, among others, will form the basis of the monitoring.
- III. **Content:** Hate speech, radical ideologies, and propaganda.

IV. **Methodology:**

A. **Multi-Agency Collaboration**

Monitoring will be carried out in partnership with relevant government agencies, civil society organizations, tech platforms, and fact-checkers. These collaborations will support efficient information sharing, coordination, and escalation of high-risk content.

B. **Capacity Building**

Stakeholders involved in monitoring will be regularly trained on digital forensics, evolving online trends, legal provisions, hate lexicon updates, psychological resilience, and ethical considerations.

C. **Use of Technology and Artificial Intelligence (AI)**

This shall involve the development of AI-powered tools and machine learning models to detect patterns, keywords, and suspicious content.

V. **Human Verification and Analysis:**

Trained analysts will manually review flagged content to assess its context, intent, and impact. This helps reduce false positives, interpret coded speech, and ensure decisions are rooted in a nuanced understanding of local dynamics.

VI. **Fact-Checking and Source Validation:**

Verified information will be cross-referenced with trusted fact-checking partners to distinguish between misinformation and legitimate opinion.

VII. **Data Analysis and Mapping:**

Trends will be analyzed across time, geography, and user demographics to detect potential hotspots and emerging threats.

7.0 Elements for Analyzing Hate speech and radical ideologies

Flagged content must be assessed systematically to determine its level of risk and potential harm. The key elements to consider are;

a) Context

Assess the environment in which the content appears—political, social, or conflict-related. Content posted during tense or sensitive periods may have a greater impact.

b) User Profile

Consider the user's influence, reach, and engagement. Look at who they are targeting, whether they operate in echo chambers, and the nature of their audience.

c) Intent

Evaluate whether the post aims to incite, mobilize, or radicalize. Intent may be explicit or hidden through coded language.

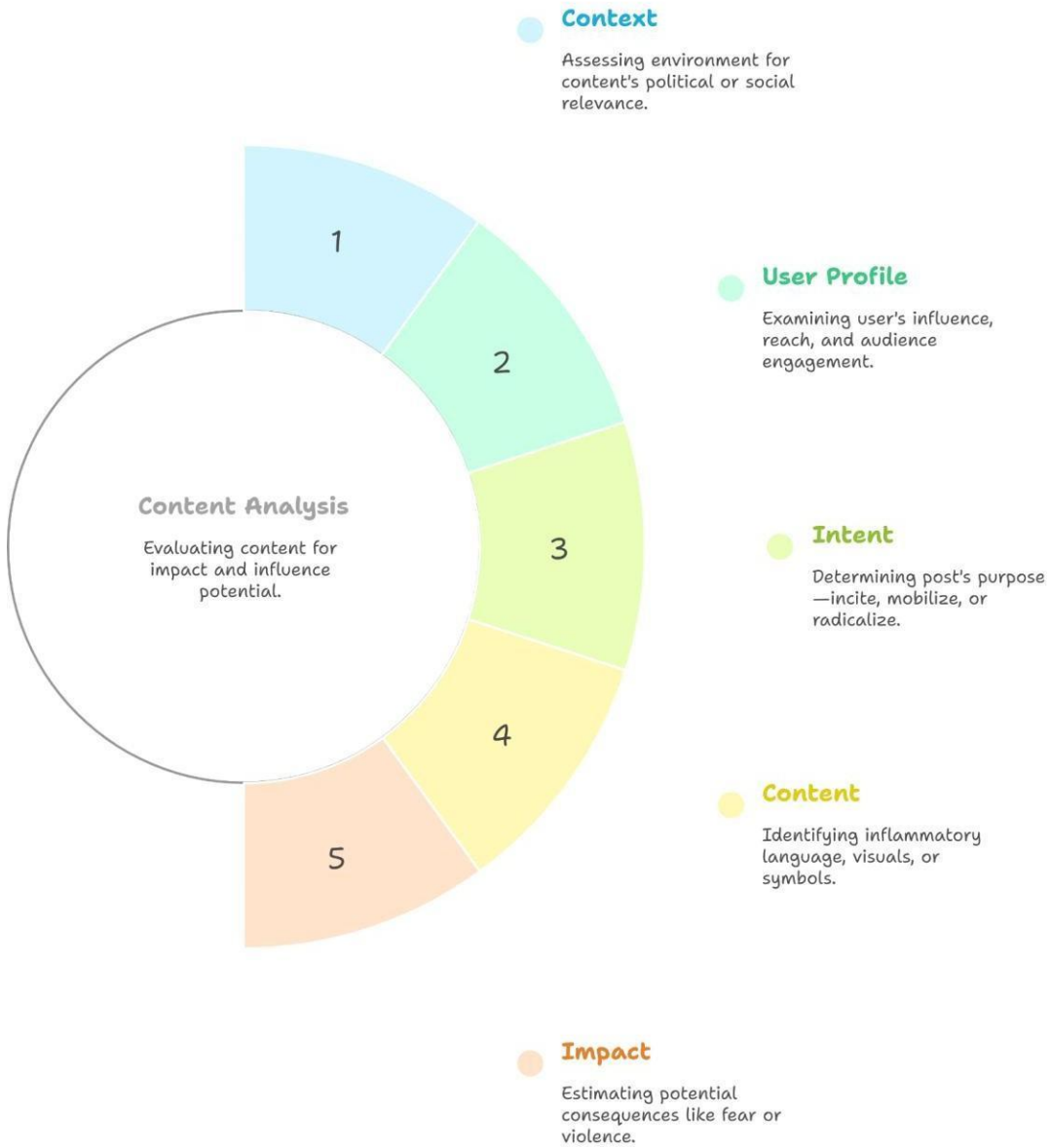
d) Content

Check for inflammatory or suggestive language, visuals, or symbols. Determine if the tone is hostile, misleading, or emotionally charged.

e) Impact

Estimate the potential consequences such as fear, harm, or social tension. Consider whether the content could lead to real-world violence or discrimination.

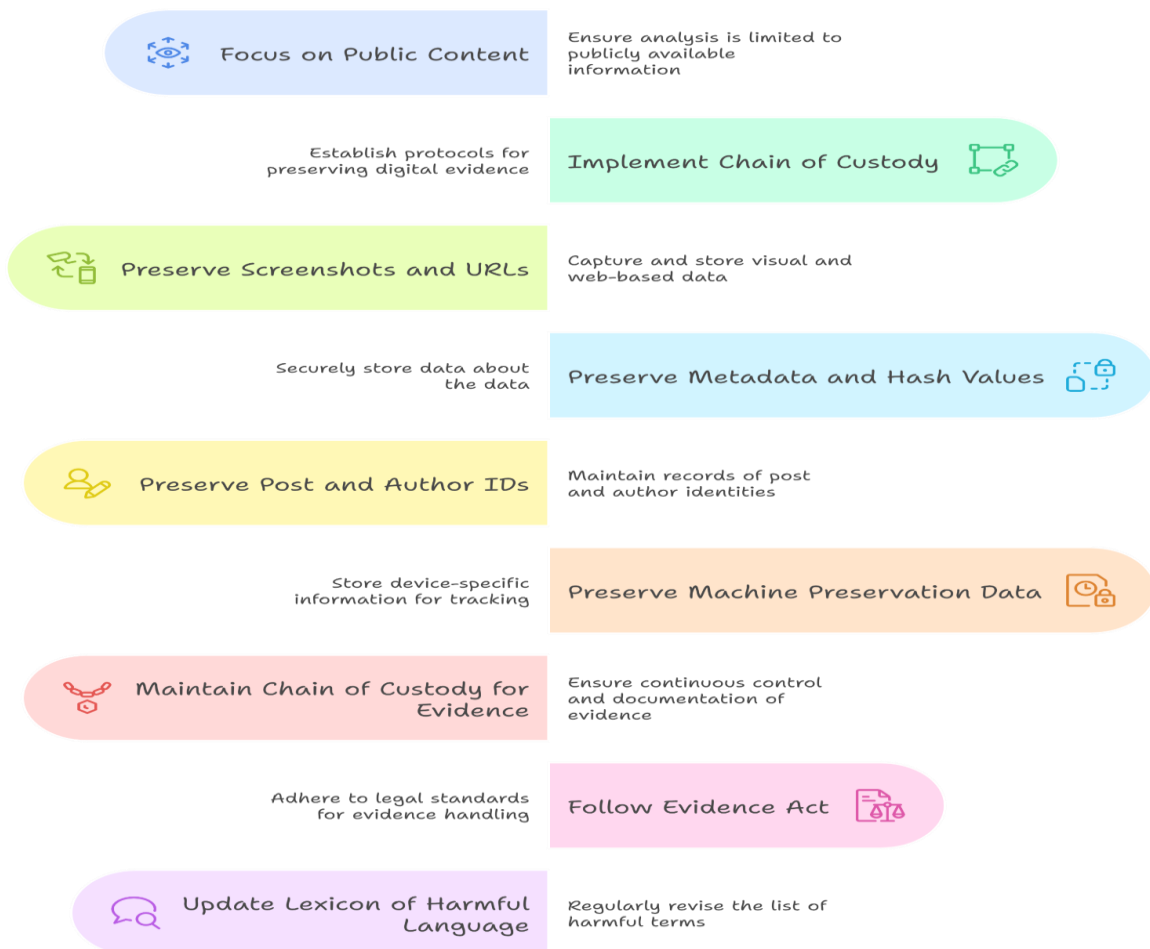
Analyzing Content Impact and Influence



8.0 Monitoring Protocols

- I. **Focus on public content:** unless legally authorized.
- II. **Implement chain of custody protocols:** Preserve screenshots, URLs, metadata, hash values, Post ID and Author ID to enable tracking of posts even if they were taken down, as well as machine preservation (IMEI, IP Address, custody) until adduced in court.
- III. **Maintain a chain of custody for electronic evidence:** follow the Evidence Act rules on admissibility of electronic evidence.
- IV. **Supervising analysts:** to update the lexicon of harmful language over time.

Digital Evidence Handling Process



9.0 Monitoring Procedures

9.1 Preparation

1. **Verify Legal Authority:** Confirm investigative jurisdiction (e.g., cybercrime division, NCIC collaboration).
2. **Obtain Monitoring Tools:** Use approved tools (e.g. Phoenix, Tweet Deck, Net Clean).
3. **Identify Keywords & Hashtags:** Tailor by context (as per the developed lexicons).

9.2 Monitoring Steps

- I. **Track public groups/pages/accounts** on digital platforms.
- II. **Apply Hate Speech Test:**
 - a. Is the language abusive, threatening, or insulting?
 - b. Is the intent to stir up ethnic, racial, or religious abuse?
 - c. Is there a risk of violence or public disorder?
- III. **Flag content using:**
 - a. Timestamps, location (if geo-tagged), user handle.
 - b. Screenshots, video/audio capture.
 - c. Digital chain of custody logs.

Steps to Initiate Cybercrime Investigation



10.0 Response Framework

- I. **Low Risk:** Issue counter-narratives, alternative messaging, counter-messaging, fact-checking, and community alerts, notify platforms, and refer for investigation, request takedown of content and bots, and continuous monitoring.
- II. **Moderate Risk:** Notify platforms and refer for investigation, request takedown of content and bots, and continuous monitoring.
- III. **High Risk:** Multi-agency approach– Escalate to relevant authority for coordinated intervention, e.g. DCI, NCTC, NCIC, notify platforms, and refer for investigation, request takedown of content and bots.
- IV. **Engage with victims at all levels:** For instance, visits to localities and offering psychosocial support
- V. **De-escalate:** to reduce risks of retaliatory hate speech and mediate underlying disputes.
- VI. **Education** and **raising awareness** between groups.
- VII. **Prosecution:** Submit well-documented cases to ODPP for hate or terrorism charges, and the investigating agency or ODPP to make requests to tech companies for user verification through court orders.
- VIII. **Rapid Response** frameworks and pathways
- IX. **Whistleblower protection**

10.1 Reporting and Oversight

- I. Publish bi-annual trend reports and assessments.
- II. Public complaint system for wrongful flagging or misconduct.

10.2 Capacity Building

- I. Train monitors and officers on:
 - a. Radicalization indicators;
 - b. Legal provisions;
 - c. Digital forensics;
 - d. Psychological resilience.

11.0 Ethical Considerations

- I. **Privacy** – protect user data and avoid mass surveillance; collect only what's necessary.
- II. **Freedom of Expression** – safeguard legitimate speech; avoid overreach or censorship.
- III. **Transparency & Accountability** – be clear about monitoring practices and allow appeals.
- IV. **Non-Discrimination** – Prevent bias in tools and human judgments; ensure fairness.
- V. **Multi-Stakeholder approach** – engagement in policy development.
- VI. **Proportionality** – ensure monitoring is risk-based, targeted, and not excessive.
- VII. **Legal Safeguards** – ground monitoring in law, with due process and oversight.

Counter-Messaging Strategy

- I. **Promote Positive Narratives** – share messages that foster unity, diversity, and peaceful coexistence.
- II. **Use Trusted Voices** – engage influencers, local leaders and community figures to deliver credible messages.
- III. **Timely and Context-Specific** – respond quickly to hate speech spikes with tailored, relevant messaging.
- IV. **Culturally and Linguistically Sensitive** – Communicate in local languages using culturally appropriate tones.
- V. **Promote awareness through digital platforms** – e.g. national sporting events like African Nations Championship (CHAN).
- VI. **Engaging Digital Formats** – use videos, memes, infographics, and relatable storytelling to reach wider audiences.
- VII. **Youth Involvement** – Involve young people through creative campaigns and user-generated peace content.

- VIII. **Platform and Fact-Checker Collaboration** – work with social media platforms and fact-checkers to counter falsehoods.
- IX. **Monitor Impact** – Track engagement and audience feedback to refine messaging.
- X. **Promote Positive Engagement** – Encourage respectful discourse and reporting of harmful content.

12.0 Safeguarding Measures

- I. **Protect Identities and Personal Data.**
- II. **Safeguard Vulnerable Groups** – Prioritize protection of minorities, children, PWDs, and abuse survivors.
- III. **Safe Reporting** – Provide secure, anonymous channels for reporting and protecting whistleblowers.
- IV. **Mental Health Support** – Offer psychosocial support and trauma-informed care for monitoring staff and victims
- V. **Ethical Use of Evidence** – use flagged content responsibly; avoid re-sharing harmful material.
- VI. **Informed Consent** – ensure voluntary and informed participation in any outreach or intervention.

Published by

National Cohesion and Integration Commission (NCIC)

Britam Towers, 17th Floor, Hospital Road, Upper Hill,

P.O. Box 7055-00100, G.P.O. Nairobi

Email: info@cohesion.go.ke

Website: www.cohesion.go.ke

ISBN: 978-0-99-702549-1

Contact Details

National Cohesion and Integration Commission (NCIC)

Britam Tower, 17th Floor, Hospital Road, Upper Hill,

P.O. Box 7055-00100, G.P.O. Nairobi

Tel: 254-20-8073264/5

Toll Free Number: SMS No. 15666

Mobile: 254-702-77700

Email: info@cohesion.go.ke

Website: www.cohesion.go.ke

X Handle: @ NCIC_Kenya

Facebook Page: National Cohesion and Integration Commission



©National Cohesion and Integration Commission

©2026

All rights reserved. This publication should not be reproduced, distributed, or transmitted in any form including photocopying or other electronic methods.