CODE FOR AFRICA

MEDA
MAPPING ELECTORAL DISINFORMATION IN AFRICA

Mapping Electoral Disinformation in Africa Toolkit

Digital Media Monitoring & Listening

Content

Content

Defending Democracy

A playbook for detecting and exposing digital subversion campaigns specifically during elections or high stakes events. It is built for journalists and CSOs tracking elections and democracy conversations as a nonpartisan resource to detect and expose influence operations by domestic and foreign state actors as well as non-state actors ranging from political activists, to paid lobbyists, conspiracists and extremist agitators.

Fighting propaganda with propaganda normalises information warfare, further polarises our societies, and erodes public trust in all information. This playbook is therefore explicitly intended for civic self-defence, and does not offer or advocate for offensive counter-measures.

Credits

The playbook is distilled from a decade of accumulated knowledge and tradecraft pioneered by Code for Africa's information integrity teams, working across 28 African countries.

- Chief strategists: **Justin Arenstein and Chris Roper**
- Editors: **Amanda Strydom, Athandiwe Saba, Doreen Wainainah, Mitchelle Awuor & Mwende Mukwanyaga**
- Lead researchers: **Jacktone Momanyi and John Ndung'u**
- Contributing researchers: **Anita Igbine (Nigeria), Bilal Tairou (Benin), CC Chargi (South Africa), Collin Kahumbi (Kenya), Christian Ngnie (Senegal), Dorcas Solonka (Kenya), Eliud Akwei (Ghana), Hanna Teshager (Ethiopia), Harriet Ogayo (Kenya), Jones Baraza (Kenya), Kúnlé Adébàjò (Nigeria), Lujain Alsedeg (Sudan), Moffin Njoroge (Kenya), Naomi Wanjiku (Kenya), Nirali Patel (Kenya), Rodgers Omondi (Kenya), Samaila Atsen Bako (Nigeria), Dr Sandra Roberts (SA), Sekamotho Ehlert (SA) , Vanessa Manessong (Cameroon),**
- Contributing editors: **Janet Heard (SA), Sisanda Ntshinga (South Africa)**
- Copyeditors: **Gloria Aradi (Kenya) and Paul Amisi (Kenya)**
- Designers: **Temidayo Oyegoke, Michael Igwe-Ebi, Bukola Onwordi**
- Project coordinators: **Virginiah Gitome, Hadeye Toure and Sarah Gowon**
- ADDO technical advisory group: **Adam Fivenson (National Endowment for Democracy), Ben Nimmo (OpenAI), Carl Miller (Centre for the Analysis of Social Media), Dani Madrid-Morales (University of Sheffield | Disinformation Research Cluster), Herman Wasserman (Centre for Information Integrity in Africa) and Nina Otte-Witte (Deutsche Welle Akademie).**

Editions

This edition of the playbook, in English, was produced with support from the MEDiA Project (Mapping Electoral Disinformation in Africa), with funding from Open Societies Foundation (OSF).

Glossary

This is a comprehensive list of terms and abbreviations used throughout the document in alphabetical order. The lexicon is internationally accepted by the International Fact-Checking Network, the European External Access Service and the United Nations.

- **ABCDE Framework** A system used to detect and label information manipulation and interference by identifying five elements: Actors, Behaviour, Content, Degree, and Effect
- **Astroturfing** A deceptive practice that involves creating a false impression of widespread, grassroots support for a particular agenda, product, or policy
- **Boolean operators** A search technique that uses specific words and symbols to expand or narrow search parameters in databases or search engines
- **Bots** Automated accounts
- **Chirpwire** A social media platform with similar features to Facebook
- **CIB** Coordinated inauthentic behaviour
- **Cyborgs** Hybrid accounts that combine automation with human oversight
- **Dark social** Content posted or shared through dark social media, which are private channels such as texts, emails, or private messages.
- **Disinformation** False or inaccurate information that is intentionally spread to mislead and manipulate people, often to make money, cause trouble or gain influence.
- **DISARM Framework** Disinformation Analysis and Risk Management framework, developed by the DISARM Foundation
- **DNS** Domain Name System, a distributed naming system for computers on the internet.
- **Domain name** A human-readable address used to access websites on the internet.
- **FIMI** Foreign information manipulation and interference.
- **IMI** I nformation manipulation and interference.
  I'lam A multi-language Islamic State media website.
- **JNIM** Jama'a Nusrat ul-Islam wa al-Muslimin, an Al-Qaeda affiliated extremist group.
- **Misinformation** False, incomplete, inaccurate/misleading information or content which is generally shared by people who do not realise that it is false or misleading.
- **Malinformation** Information that is based on truth (though it may be exaggerated or presented out of context) but is shared with the intent to attack an idea, individual, organisation, group, country or other entity.

Glossary

- **OSINT** (Open Source Intelligence) is the collection, processing, and analysis of publicly available information to produce actionable intelligence.
- **PIP** Politically Influential Person is a person who operates or has operated within the past year in an official position.
- **Synthetic media** digital content, including images, video, audio, and text, that is partially or fully generated using artificial intelligence (AI) or machine learning
- **SOCMINT** (Social Media Intelligence) is a sub-branch of open-source intelligence (OSINT) that involves the collection, analysis, and interpretation of data from social media platforms to gain actionable insights
- **Trolls** People who intentionally provoke or disrupt online discussions to amplify misleading narratives
- **TTP** Tactics, techniques, and procedures.
- **Web3** a decentralised internet in which users have more control over their data and interactions through blockchain technology.

1.

Introduction

Introduction

Covert manipulators increasingly subvert democratic processes, using sophisticated digital techniques to deceive, polarise, and incite the public for political gain. Across Africa, illicit influence operations have become a common tactic in hybrid warfare, undermining trust in government institutions and destabilising nations.

State-affiliated 'StratCom' professionals, ideological campaigners, and freelance 'keyboard warriors' drive these operations, working for a range of clients. Digital mercenaries rapidly adapt their tools and techniques to evade civic defenders.

This playbook equips human rights defenders with early warning systems to detect and expose illicit influence campaigns. It introduces key aspects of information disorder and resilience, highlighting strategies to safeguard information integrity, support free speech, and combat information pollution and organised crime.

Users will learn how to map information ecosystems, monitor key actors, identify threats, and track content to protect the integrity of information flows. The playbook explores manipulation tactics, including disinformation, coordinated inauthentic behaviour, hate speech, and suppression, using case studies from Africa and beyond to illustrate these challenges. It also examines how digital platforms address online harm and how to counter these threats through tools like social media intelligence (SOCMINT), human intelligence (HUMINT), network mapping, and behavioural analysis.

A key focus is on building an early warning system to disrupt threats. The document outlines methods for media mapping, content monitoring, and tracking information trends to identify emerging risks. It also provides strategies for debunking misinformation, countering myths, and moderating harmful content. CivicSignal's MediaData plays a critical role in mapping media influence by profiling news organisations, media owners, regulators, and industry professionals. MediaCloud, in turn, allows users to collect, analyse, and visualise news stories, offering a real-time picture of how information spreads across the continent. Together, these tools help defenders track shifts in public sentiment, identify threats to electoral integrity, and expose influence operations designed to manipulate public opinion.

Practical guidance on team structure, workflow, and security ensures that those working in information resilience operate efficiently and safely. Secure communication, information management, and the amplification of insights are essential to driving meaningful action in the fight for information integrity.

A regularly updated online toolkit complements this playbook, offering the latest resources for strengthening information resilience. A network of rapid-response investigators from the **African Digital Democracy Observatory (ADDO)** and the **African Fact Checking Alliance (AFCA)** supports these efforts, alongside operational security assistance from the TrustLab.

This playbook serves as a foundational resource. Civic defenders seeking to enhance their skills can access modular, step-by-step training through the ADDO Academy.

2.

Information resilience: Why is it important?

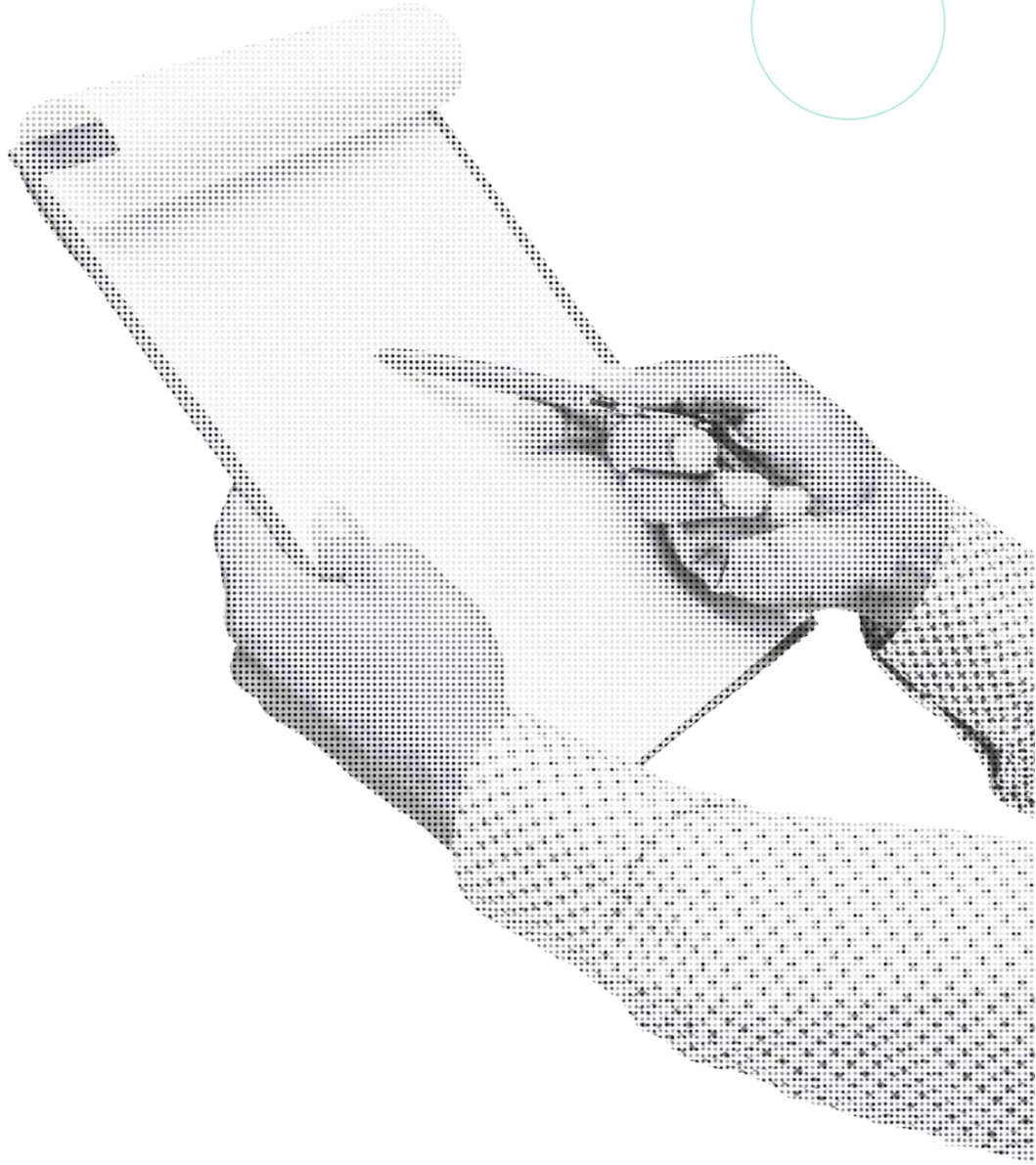
2. Why is it important?

This section will introduce you to information resilience, information integrity, and information pollution. It will discuss the importance of information resilience, what threatens it, and how it can be protected. It will also highlight some of the measures social media platforms are taking to safeguard information integrity.

Information resilience refers to the ability to protect information integrity, while managing threats and human rights.

Information integrity ensures digital content is factual and trustworthy. It requires clear sources to ensure transparency and credibility; verifiable factual accuracy; full context to prevent misinterpretation or manipulation; and true representation which keeps content unaltered.

However, this needs to be balanced with the freedom of speech and association, and the right to dissent. Article 19 of the UN's Universal Declaration of Human Rights affirms that, 'everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference, and to seek, receive and impart information and ideas through any media and regardless of frontiers'. These freedoms allow individuals to engage in public discourse, exchange ideas, and contribute to transparency and accountability in governance and democratic processes such as elections.



Safeguarding Information Integrity

Safeguarding information integrity also requires balancing these freedoms with measures to counter harmful content, manipulation, and threats to social cohesion. The core dimensions of freedom of expression in information spaces are as follows:

Protected speech vs harmful content

Protected speech versus harmful content, which distinguishes free speech with speech that incites violence, fuels dangerous falsehoods, or spreads hate.

Platforms and amplification

The right to free speech also encompasses who gets amplified, how, and the impact of the amplification on information integrity.

Counter-speech & refutation

Dissent allows for the challenge of misinformation, the sharing of alternative perspectives, and the holding of powerful actors accountable.

Transparency in restrictions

Any limits on speech must be clear, consistently applied, and proportionate, regardless of political status.

Association and collective action

Freedom of association enables civil society groups, communities, and fact-checkers to tackle information manipulation together.

For the purpose of this playbook, the focus is on digital threats to information integrity, how to monitor it, how platforms attempt to manage these, and how to counter the threats.

Information integrity faces systematic attacks on digital spaces in five ways:

- **Coordinated inauthentic behaviour (CIB):**
  Networks of accounts working together to artificially amplify narratives.
- **Cross-platform manipulation:**
  Synchronised campaigns leveraging multiple platforms to reach diverse audiences.
- **Artificial engagement:**
  Bot-driven interactions creating false impressions of content popularity or consensus.
- **Synthetic media:**
  Content generated using artificial intelligence (AI), designed to fabricate photos, videos or wording.
- **Strategic timing:**
  The deployment of false narratives at critical moments in electoral or governance processes.

To counter these threats, experts have developed various frameworks to strengthen information integrity:

- **Verification:**
  Structured approaches to confirm whether information aligns with authenticated reality.
- **Transparency:**
  Systems that trace the origins and journey of information across digital spaces.
- **Context integrity:**
  Methods to detect the manipulation of information through selective presentation or de-contextualisation by analysing missing elements that could alter interpretation.
- **Authenticity:**
  Tools to ensure content remains as originally created, identifying manipulations such as deepfakes, selective edits, and synthetic media.
- **Integration:**
  Cohesive systems that combine the aforementioned approaches for a more comprehensive response.

Social media platforms have integrated some levels of moderation approaches to balance freedom of expression with information integrity.

Content labelling

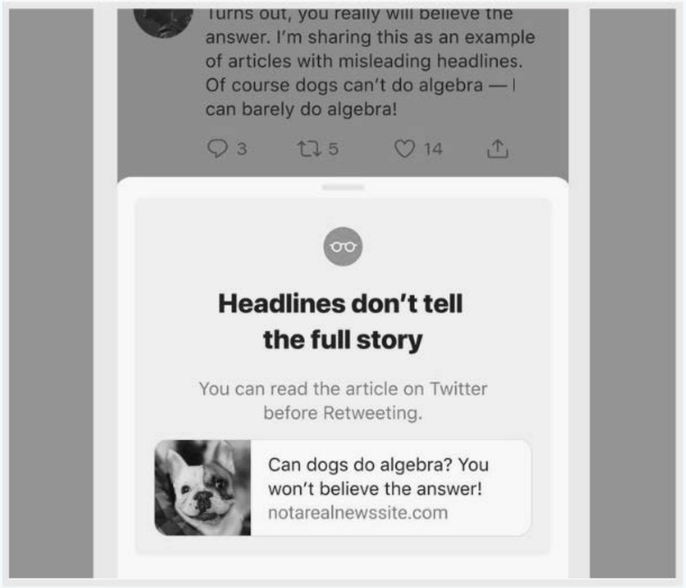
Content labelling on social media platforms involves attaching visual and/or textual information to user-generated content to provide context for viewers, serving as 'information about information' to help users better assess what they encounter.

These labels can direct users to mainstream media links by providing additional context or fact-checks, helping them access more reliable sources and make informed decisions.

Social media platforms flag and label various types of content, including misinformation, disputed information, conspiracy theories, misleading content, AI-generated content, manipulated media, content that could cause confusion about products or services, and content deliberately using AI. Platforms may use labels to indicate disputed messages, provide context, or inform users about AI-created content.

X 'read before reposting' prompt

This prompt introduces informed sharing by encouraging users to click on the article before amplifying it. This approach slows impulsive sharing without restricting content, balancing information integrity with free expression.



A mock-up graphic of the prompt displayed when an X user attempts to repost a post with an article **(Source: CfA via Coywolf)**

Meta's AI content labelling

Meta adds 'AI info' labels to video, audio, and image content that their systems detect as AI-generated or when users self-disclose AI creation. When content is determined to create 'a particularly high risk of materially deceiving the public on a matter of importance', Meta may apply more prominent labels to provide users with additional context. The system works alongside their fact-checking network, which reviews potentially misleading AI-generated content.

X community notes

Meta adds 'AI info' labels to video, audio, and image content that their systems detect as AI-generated or when users self-disclose AI creation. When content is determined to create 'a particularly high risk of materially deceiving the public on a matter of importance', Meta may apply more prominent labels to provide users with additional context. The system works alongside their fact-checking network, which reviews potentially misleading AI-generated content.

A mock-up graphic of the notification an X user gets when a community note is added to a post they interacted with (Source: CfA via X)

Algorithmic demotion

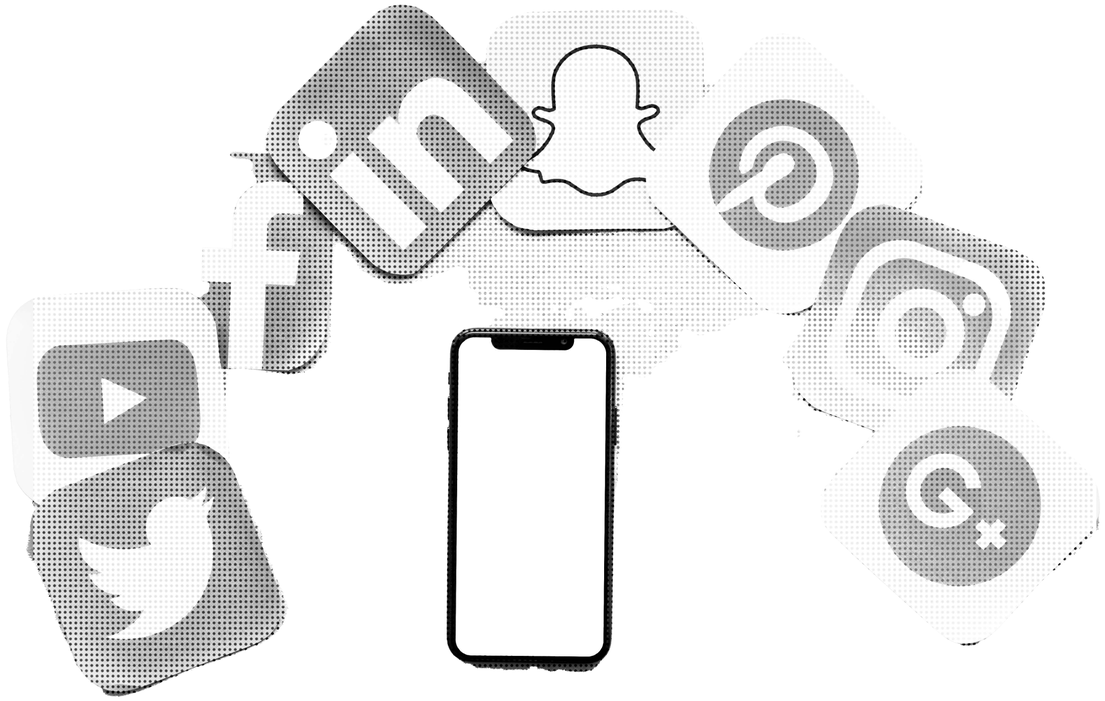
This is when a platform's algorithm automatically reduces the visibility of a post, causing it to appear less frequently in users' feeds. It often happens due to factors such as flagged violations.

YouTube's reduced recommendations approach

YouTube's approach to handling borderline content— content that does not explicitly violate the platform's policies, includes reducing their algorithmic reach instead of removing them. This preserves users' right to post while addressing concerns about promoting potentially harmful material.

Facebook's demotion approach

Facebook's algorithmic demotion reduces the visibility of content deemed problematic or low-quality, such as misinformation verified by fact-checkers, click-bait, and posts from users who frequently violate the platform's policies. The extent of demotion varies based on factors such as the user's violation history and the system's confidence in the assessment. Additionally, Facebook employs personalised ranking to limit the reach of 'borderline' content that nears policy violation thresholds.



2.1 Information pollution

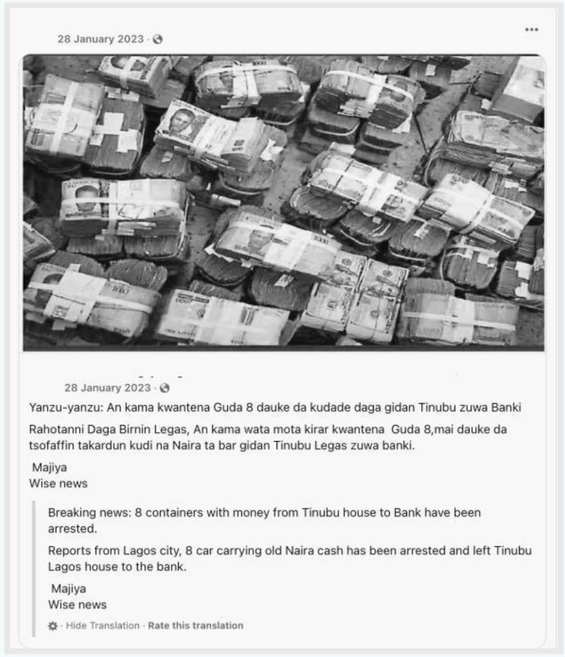
Information pollution refers to unwanted, irrelevant, or deceptive content, including misinformation, disinformation, malinformation, and manipulated content. It weakens someone's ability to detect manipulated information, especially during polarising events such as elections.

One digital phenomenon which contributes to information pollution is **disinformation as organised crime** - a shadowy 'disinformation-for-hire' industry where paid influencers, influencer companies, and propagandists offer services to spread falsehoods and manipulate public perception. These operations deploy tactics such as smear campaigns, deepfake content, and the coordinated amplification of false narratives to undermine opponents, sway elections, and destabilise societies.

Real-life example
Disinformation-for-hire groups during Nigeria's 2023 general elections

EXAMPLE

During the 2023 Nigerian elections, a BBC investigation exposed disinformation-for-hire groups manipulating public perception. Major political parties hired influencers with large followings on platforms such as Facebook and X to fabricate false stories about opponents.



A screenshot of false information a micro-influencer posted during the Nigerian elections (Source: Facebook)

3.

Information ecosystems: Finding your way around

3. **Information ecosystems** Finding your way around

This segment explores the information ecosystem, its resilience, and assessment methods, including mapping key attributes. It distinguishes between open and dark social media, explains how to track influential actors, and guides lexicon development for monitoring online activity. Additionally, it introduces frameworks like ABCDE and D-RAIL for deeper analysis.

An information ecosystem refers to how communities interact within information and communication systems. It encompasses the channels through which information circulates, influences decision-making, and shapes public discourse.

During elections, understanding this ecosystem is essential to assessing how information, whether accurate, manipulative or misleading, impacts public opinion and voter behaviour. The ecosystem includes channels such as messaging applications, social media and traditional media, which play distinct roles in shaping election conversations.

A resilient information ecosystem is;

a. **Reflective,** continuously evolving to adapt to uncertainty and change by modifying standards and norms based on emerging evidence.
b. **Robust** with well-designed and managed physical structures that can withstand the impacts of nuances without significant damage or loss of function.
c. **Redundant t**o strengthen itself, providing spare capacity and diverse solutions to accommodate disruptions or surges in demand.
d. **Flexible** to allow for seamless adaptation, integrating new technologies and traditional knowledge to enhance resilience.
e. **Resourceful** to ensure that people and institutions can quickly find alternative ways to achieve their goals, mobilising resources effectively even under stress.
f. **Inclusive** to foster broad participation, ensuring all voices, especially the most vulnerable are heard and considered.
g. **Integrated** to align its different components promoting consistency, collaboration and efficiency.

3.1. Information ecosystem assessments

Assessing the information ecosystem involves understanding how people and communities find, share, value, and trust information in their local contexts, whether from the media or other sources. Information can spread through various channels, such as word of mouth, community leaders, local media, social media, and other means.

3.1.1. Mapping the information economy

The information economy focuses on the value derived from the creation, consumption, and exchange of information. It consists of three main pillars: **attention, influence, and monetisation.**

Attention

Attention refers to the focus and time individuals or groups devote to content, such as news, political Fements, or social media posts. As the currency of the information economy, attention drives engagement, reach, and virality.

How to evaluate attention during an election campaign

Content virality determines how quickly and widely political content spreads. Certain posts, videos, or memes can gain momentum when they capture public interest, trigger strong emotions, or tie into ongoing political conversations.

Example: A political candidate's video might accumulate millions of views overnight, or a meme about a controversial debate could go viral on X due to mass reposting.

Questions to consider:

- What type of content is being most shared? (e.g., memes, news, or videos)
- What content is being most shared?
- Is there a specific trigger event causing viral attention?

Engagement reflects how audiences interact with political content, through likes, shares, comments and replies. High engagement levels can indicate strong interest but can also signal coordinated efforts to amplify certain messages artificially.

Example: A political candidate's Facebook post might receive thousands of comments, or a trending hashtag on X could spark widespread discussions.

Questions to consider:

- What is the level of engagement (comments, likes, or shares)?
- Are engagement patterns similar across different demographics or regions?
- Are there signs of coordinated efforts, such as bot activity, to amplify engagement?

Content saturation refers to how dominant certain content becomes within the media landscape. When a topic or message appears repeatedly across multiple platforms, it can shape public discourse by overshadowing other issues.

Example: When a political scandal is covered by every major news site, reposted across social media, and discussed in various formats, including memes, ads, and blog posts.

Questions to consider:

- Is the content appearing frequently on different platforms or news sources?
- How repetitive is the content, and does it seem to dominate certain spaces?
- Are there signs of content overload to overwhelm or manipulate public perception?

Impact on public opinion. The level of attention a topic or candidate receives can directly influence public opinion. When certain narratives dominate media coverage, they shape how people think, feel, and ultimately vote.

Example: Continuous negative press about a candidate may lead to a decline in public support, while positive messaging from another candidate could bolster voter confidence.

Questions to consider:

- How is attention on specific topics influencing public opinion?
- Is attention on certain issues skewing the public's perception of candidates or political parties?
- What is the narrative emerging from the attention on specific issues?

Suspicious attention patterns. Identifying unnatural spikes in attention helps detect potential disinformation or influence campaigns. While organic attention reflects genuine public interest, it can also be artificially manipulated through coordinated efforts, such as bot networks, fake accounts, and paid influencers.

Example: A sudden surge of interest in a particular topic, without any clear trigger, could indicate engineered amplification. Similarly, if multiple social media accounts post identical messages simultaneously, it may suggest an orchestrated effort to push a narrative.

Questions to consider:

- Are there sudden spikes in attention that seem unnatural or engineered?
- Do we see coordinated content across multiple accounts or platforms?
- Is there evidence of bots or fake accounts amplifying attention?

Example:
Clickbait to trick audience during US elections

EXAMPLE

In the run-up to the 2024 US elections, deceptive political ads flooded Facebook and Instagram. Advertising networks such as Patriot Democracy and Kontrol LLC ran over 160,000 ads across 340 Facebook pages in English and Spanish. These ads misled users, promising free government money or health insurance, and reached 38 million views. Once clicked, they redirected victims to unethical insurance agents, leading to unauthorised charges and lost health coverage. Despite Meta removing some ads, weak enforcement allowed new ones to keep circulating.

This case shows how attention in the information economy drives engagement and spreads harmful content when oversight is inconsistent.

Two ads run by Patriot Democracy falsely promised government subsidy checks (Source: ProPublica)

Influence

shapes behaviour, beliefs, and emotions by spreading information strategically. Emotional appeal, framing, and narratives play roles in shaping public opinion and voter behaviour, with social media amplifying messages, especially during elections.

How to evaluate influence during an election campaign

i. **Narratives** shape how people perceive events, influencing opinions and driving actions.

Example: A story may present a candidate as the 'saviour' of the nation or paint them as corrupt. Media coverage can frame an event as either a victory or a scandal, depending on political affiliation.

Questions to consider:

- What narratives are emerging in the election cycle?
- How does the media frame political events and candidates?
- Do certain narratives receive more amplification while others are ignored?

ii. **Framing** presents information in a way that highlights certain aspects while downplaying others, shaping interpretation and emotional response.

Example: A political advert may emphasise a candidate's achievements, portraying them as a strong leader. A news headline might describe a protest as 'civil unrest' or a 'peaceful demonstration'.

Questions to consider:

- How does framing influence voters' perceptions?
- Which facts or aspects receive emphasis to guide interpretation?
- How does framing shape public understanding of events or candidates?

iii. **Emotional appeal** triggers strong emotions—anger, fear, hope, or pride—to gain support, influence behaviour, or shape decisions.

Example: A campaign advert may show vulnerable people to evoke empathy for a political cause. A speech might use fear of future threats to rally support.

Questions to consider:

- Which emotions does the content target (e.g. anger, fear, hope)?
- How does the content provoke a specific emotional response?
- Does emotional appeal overshadow factual discussion or rational debate?
-

iv. **Call to action** prompts specific behaviour, mobilising supporters, persuading audiences, or shifting voting patterns.

Example: A political advert may encourage voting by using persuasive tactics such as social proof. A social media post may urge people to 'take action' by attending an event or sharing the post.

Questions to consider:

- Does the content seek to change voter behaviour, such as encouraging activism, donations, or turnout?
- What types of calls to action appear in the messaging?
- Does the content use social pressure or groupthink to manipulate behaviour?

v. **Mis- and disinformation** distort public opinion by spreading misleading or false information.

Example: An X post may claim false election results to dissuade voters or spread doubt. A fabricated news story may falsely accuse a candidate of a crime to damage their reputation.

Questions to consider:

- Does the shared information appear accurate, or does it seem misleading or false?
- Do coordinated efforts, such as fake news websites or bots, amplify disinformation?
- How does false information frame issues to sway opinions?

vi. **Echo chambers** reinforce existing beliefs by exposing individuals to repeated messages while limiting opposing views.

Example: A voter may only see content from one political perspective, strengthening their biases. A political group may create a private space that circulates content solely supporting its stance.

Questions to consider:

- Do echo chambers form around certain political messages or candidates?
- How isolated are different groups from diverse political perspectives?
- Does content circulate in closed circles with little exposure to opposing viewpoints?

Real-life example:

Algorithms shape vote perception during German elections

EXAMPLE

Ahead of the 2025 German federal elections, an investigation by Global Witness, an international non-profit research tank, exposed how social media algorithms shaped voter perceptions. Non-partisan users on Instagram, TikTok, and X were disproportionately fed right-leaning content, with 78% of TikTok's recommendations and 64% of X's favouring the far-right party Alternative für Deutschland (AfD).

By amplifying anti-immigration and nationalist themes, these platforms framed political discourse, evoking anger, fear, and pride to influence engagement. The constant exposure created digital echo chambers, reinforcing biases and mobilising AfD supporters while limiting alternative viewpoints.

Click here to read the Global Witness article.

Monetisation

Monetisation refers to generating income from viral content, often through political ads, misinformation, or narratives that drive engagement and reach. Monetisation within the election ecosystem operates through multiple channels, including:

i. **Political ads** are paid advertisements designed to spread specific political messaging and influence voters, often targeting particular demographics.

Example: A political campaign may run ads on Facebook aimed at swing voters in regions. On Instagram, ads may highlight a candidate's policy positions and accomplishments.

Questions to consider:
Who is funding the ads?
What messages are being pushed through paid advertisements?
How targeted are these ads, and which demographics are being reached?

ii. **Influencer endorsements** involve social media personalities promoting political views or candidates, often through paid partnerships, shaping public opinion by leveraging their follower base.

Example: A popular influencer may endorse a political candidate in exchange for financial compensation. An influencer might also post about political issues using a sponsored hashtag to promote a specific agenda.

Questions to consider:
- Are influencers being paid to endorse political candidates or messages?
- How transparent are these endorsements, and is there a financial incentive involved?
- How much influence do these figures have on their audiences?

iii. **Troll farms** are organised groups that create and amplify misleading or false content to manipulate public opinion, often for financial or political gain. They use bots, fake accounts, and paid workers to spread their messages.

Example: A troll farm may spread fake news about an election candidate through thousands of fake accounts. Paid operatives might push divisive political content on social media to manipulate voter sentiment.

Questions to consider:
- Are coordinated, fake accounts spreading content for profit?
- How are troll farms financially benefiting from their activities?
- What content are troll farms focusing on, and how is it impacting public discourse or voting behaviour?

iv. **Bots and automated engagement** use artificial intelligence to generate fake interactions – such as comments, likes, or shares – to increase the visibility of political messages and boost monetisation.

Example: Bots may amplify a political hashtag, making it trend and increasing ad revenue for those promoting it. Automated systems can also generate fake comments that enhance engagement and attract advertisers.

Questions to consider:
- Are bots or automated systems involved in generating false engagement?
- How is this artificial engagement influencing monetisation (e.g. ad revenue or political impact)?
- Who is behind the creation and operation of these automated systems?

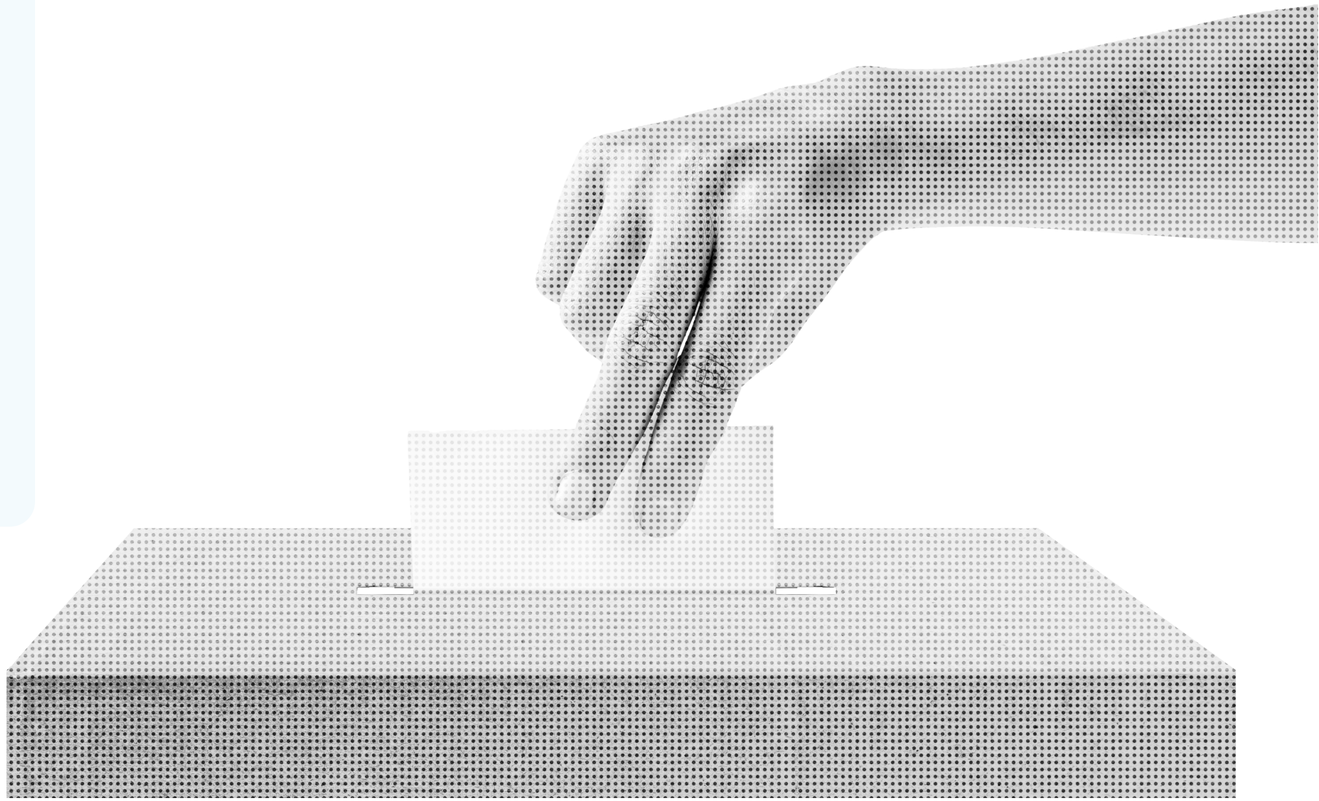
Real-life example:
2024 Romanian presidential elections

EXAMPLE

During the 2024 Romanian presidential elections, pro-Russia nationalist Călin Georgescu won the first round, propelled by a coordinated TikTok campaign. Investigations revealed that third-party platforms paid influencers to spread hashtags, attracting bots and amplifying his online presence. Authorities responded with tax probes into influencers and their financial backers, focusing on €1million allegedly funnelled by businessman Bogdan Peșchir.

Amid concerns over voter manipulation, Romania's constitutional court annulled the election, prompting some influencers to flee. The ruling coalition rescheduled the vote for May 2025, with ongoing investigations into financial transactions, influencer roles, and potential Russian interference.

Click here to read the Deutsche Welle, Politico, and Radio Moldova articles on this topic.



3.1.2. What are ‘open’ and ‘dark’ platforms?

Social media platforms enable users to communicate and share content through images, links, text, and videos. Major platforms include Facebook, Instagram, Rumble, Telegram, TikTok, TruthSocial, VKontakte (VK), WhatsApp, X, and YouTube.

Social media sharing happens in two main spaces:

Open social media

These are publicly accessible platforms where content is visible to anyone and can be tracked using open source intelligence (OSINT) tools. On open platforms such as X and public Facebook groups and pages, conversations are visible, allowing campaigners, journalists, and researchers to track engagement, mis-/disinformation, and political sentiment in real time. Public debates, trends, and viral posts shape narratives that influence voter behaviour.

Dark social media

These are encrypted or private spaces where conversations remain hidden from public view, making monitoring difficult. Platforms such as Telegram, WhatsApp, and private Facebook groups enable discussions that evade scrutiny, allowing mis-/disinformation and manipulation campaigns to spread unchecked. These closed environments often function as echo chambers, reinforcing specific narratives without external challenge.

3.1.3. Identifying the actors

Tracking politically influential persons (PIPs), entities, and parties before elections helps detect disinformation campaigns, monitor narratives, and flag harmful content.

Early disinformation detection: Identifies false narratives before they spread.
Social media monitoring: Flags manipulative or harmful content in real time.
Narrative tracking: Alerts on shifts in rhetoric and political allegiances.

Mapping PIPs across platforms reveals their influence, role in shaping public discourse, and shifts in their stances. This process, managed through tools such as AirTable, Excel, or Google Sheets, requires a strict data schema for consistency.

Below is the data schema of how one should map social media profiles.

id	name_id	full_name	first_name	middle_name	last_name	gender	party_id	Political Party Name	Abbrv	office
pers01	radjabho_tebabho__soborabo	Radjabho Tebabho Soborabo	Soborabo	Tebabho	Radjabho	M	pty007	CONGOLAIS UNIS POUR LE CHAN	CUC	Presidential
pers02	mutamba_tungunga_constant	Mutamba Tungunga Constant	Constant	Tungunga	Mutamba	M	pty028	DYNAMIQUE PROGRESSISTES RE	DYPRO	Presidential
pers03	katumbi_chapwe_moise	Katumbi Chapwe Moise	Moise	Chapwe	Katumbi	M	pty004	ENSEMBLE POUR LA REPUBLIQU	ENSEMBLE	Presidential
pers04	sesanga_hipungu_dja_delly	Sesanga Hipungu Dja Delly	Delly	Hipungu Dja	Sesanga	M	pty010	PARTI DE L'ENVOL DE LA RD.CON	ENVOL	Presidential
pers05	anzuluni_isiloketshi_floribert	Anzuluni Isiloketshi Floribert	Floribert	Isiloketshi	Anzuluni	M	pty001	Independent	IND	Presidential
pers06	baende_etafe_eliko_jean_claude	Baende Etafe Eliko Jean Claude	Jean Claude	Etafe Eliko	Baende	M	pty001	Independent	IND	Presidential
pers07	bolamba_tony_cassius	Bolamba Tony Cassius	Tony Cassius	Tony	Bolamba	M	pty001	Independent	IND	Presidential
pers08	buse_falay_georges	Buse Falay Georges	Georges	Falay	Buse	M	pty001	Independent	IND	Presidential
pers09	ikofu_mputa_mpunga_marie_josse	Ikofu Mputa Mpunga Marie-Josse	Marie-Josse	Mputa Mpunga	Ikofu	F	pty001	Independent	IND	Presidential
pers10	kazadi_kanda_rex	Kazadi Kanda Rex	Rex	Kanda	Kazadi	M	pty001	Independent	IND	Presidential
pers11	kikuni_masudi_seth	Kikuni Masudi Seth	Seth	Masudi	Kikuni	M	pty001	Independent	IND	Presidential
pers12	majondo_mwamba_patrice	Majondo Mwamba Patrice	Patrice	Mwamba	Majondo	M	pty001	Independent	IND	Presidential
pers13	masalu_anedu_andre	Masalu Anedu Andre	Andre	Anedu	Masalu	M	pty001	Independent	IND	Presidential
pers14	mudekereza_bisimwa_justin	Mudekereza Bisimwa Justin	Justin	Bisimwa	Mudekereza	M	pty001	Independent	IND	Presidential
pers15	mukwege_mukengere_denis	Mukwege Mukengere Denis	Denis	Mukengere	Mukwege	M	pty001	Independent	IND	Presidential
pers16	ngalasi_kurisini_aggrey	Ngalasi Kurisini Aggrey	Aggrey	Kurisini	Ngalasi	M	pty001	Independent	IND	Presidential
pers17	ngoy_ilunga_wa_theodore	Ngoy Ilunga Wa Theodore	Theodore	Ilunga Wa	Ngoy	M	pty001	Independent	IND	Presidential
pers18	nkema_liloo_bokonzi_loli	Nkema Liloo Bokonzi Loli	Loli	Liloo Bokonzi	Nkema	M	pty001	Independent	IND	Presidential
pers19	tshiani_k_musadiamvita_noel	Tshiani K Musadiamvita Noel	Noel	Musadiamvita	Tshiani K	M	pty001	Independent	IND	Presidential
pers20	tshisekedi_tshilombo_felix_antoine	Tshisekedi Tshilombo Felix Antoine	Felix Antoine	Tshilombo	Tshisekedi	M	pty001	Independent	IND	Presidential
pers21	fayalu_madidi_martin	Fayalu Madidi Martin	Martin	Madidi	Fayalu	M	pty002	Lamuka Fayulu	LAMUKAFAYULU	Presidential
pers22	matata_ponyo_mapon	Matata Ponyo Mapon	Mapon	Ponyo	Matata	M	pty009	LEADERSHIP ET GOUVERNANCE	LGD	Presidential
pers23	diongo_shamba_franck	Diongo Shamba Franck	Franck	Shamba	Diongo	M	pty003	Mouvement Lumumbiste Progressist	MLP	Presidential
pers24	muzito_fumutshi_adolphe	Muzito Fumutshi Adolphe	Adolphe	Fumutshi	Muzito	M	pty089	Nouvel Élan	NOU.EL	Presidential

A screenshot showing the data schema for mapping candidates in the 2023 Democratic Republic of the Congo (DRC) presidential election (Source: CfA using PIPs data)

id	name_id	full_name	note_fb	fb_url	note_ig	instagram_url	note_tw	twitter_url	note_tk
pers01	radjabho_tebabho__soborabo	Radjabho Tebabho Soborabo	N/A	N/A	N/A	N.A	official	https://twitter.com/soborabo	N/A
pers02	mutamba_tungunga_constant	Mutamba Tungunga Constant	campai...	https://www.facebook.com/ConstantMutambaC	campaign	https://www.instagram.com/constantmutamb	official	https://twitter.com/ConstantMutamba	official
pers03	katumbi_chapwe_moise	Katumbi Chapwe Moise	campai...	https://www.facebook.com/mkatumbi	campaign	https://www.instagram.com/moise_katumbi_	official	https://twitter.com/moise_katumbi	N/A
pers04	sesanga_hipungu_dja_delly	Sesanga Hipungu Dja Delly	campai...	https://www.facebook.com/d.sesanga	campaign	https://www.instagram.com/dsesanga2023/	official	https://twitter.com/DSESANGA	official
pers05	anzuluni_isiloketshi_floribert	Anzuluni Isiloketshi Floribert	campai...	https://www.facebook.com/profile.php?id=1000	N/A	N.A	campaign	https://twitter.com/FloribertAnzu	N/A
pers06	baende_etafe_eliko_jean_claude	Baende Etafe Eliko Jean Claude	campai...	https://www.facebook.com/Etafeliko	N/A	N.A	campaign	https://twitter.com/JC_Baende	N/A
pers07	bolamba_tony_cassius	Bolamba Tony Cassius	campai...	https://www.facebook.com/officieltonybolamba	campaign	https://www.instagram.com/tonycassiusbolar	campaign	https://twitter.com/TonyBolamba	N/A
pers08	buse_falay_georges	Buse Falay Georges	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers09	ikofu_mputa_mpunga_marie_josse	Ikofu Mputa Mpunga Marie-Josse	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers10	kazadi_kanda_rex	Kazadi Kanda Rex	campai...	https://www.facebook.com/KAZADIREX	campaign	https://www.instagram.com/rex_kazadi/	campaign	https://twitter.com/kazadi_rex	N/A
pers11	kikuni_masudi_seth	Kikuni Masudi Seth	campai...	https://www.facebook.com/sethkikuni	N/A	N.A	campaign	https://twitter.com/sethkikuni	N/A
pers12	majondo_mwamba_patrice	Majondo Mwamba Patrice	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers13	masalu_anedu_andre	Masalu Anedu Andre	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers14	mudekereza_bisimwa_justin	Mudekereza Bisimwa Justin	campai...	https://www.facebook.com/justin.mudekereza5	N/A	N.A	campaign	https://twitter.com/JustinMudek5	N/A
pers15	mukwege_mukengere_denis	Mukwege Mukengere Denis	campai...	https://www.facebook.com/DrDenisMukwege	personal	https://www.instagram.com/drdenismukwege	campaign	https://twitter.com/DenisMukwege	N/A
pers16	ngalasi_kurisini_aggrey	Ngalasi Kurisini Aggrey	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers17	ngoy_ilunga_wa_theodore	Ngoy Ilunga Wa Theodore	N/A	N/A	N/A	N.A	N/A	N/A	N/A
pers18	nkema_liloo_bokonzi_loli	Nkema Liloo Bokonzi Loli	campai...	https://www.facebook.com/Changefordrc	personal	https://www.instagram.com/lolinkema/	campaign	https://twitter.com/LoliNkema	N/A
pers19	tshiani_k_musadiamvita_noel	Tshiani K Musadiamvita Noel	campai...	https://www.facebook.com/Tshianiforpresident	campaign	https://www.instagram.com/noelktshiani/	campaign	https://twitter.com/NoelKTshiani	N/A
pers20	tshisekedi_tshilombo_felix_antoine	Tshisekedi Tshilombo Felix Antoine	campai...	https://www.facebook.com/PresidentFelixTshise	N/A	N.A	campaign	https://twitter.com/FelixUdps	N/A
pers21	fayalu_madidi_martin	Fayalu Madidi Martin	campai...	https://www.facebook.com/martinfayuluofficiel	N/A	N.A	campaign	https://twitter.com/MartinFayulu	N/A
pers22	matata_ponyo_mapon	Matata Ponyo Mapon	campai...	https://www.facebook.com/matataponyomapon	campaign	https://www.instagram.com/matataponyomap	campaign	https://twitter.com/Mapon_Matata	campaign
pers23	diongo_shamba_franck	Diongo Shamba Franck	campai...	https://www.facebook.com/profile.php?id=1000	campaign	https://www.instagram.com/mlprdc/	campaign	https://twitter.com/mlprdc	campaign
pers24	muzito_fumutshi_adolphe	Muzito Fumutshi Adolphe	campai...	https://www.facebook.com/profile.php?id=1000	N/A	N.A	campaign	https://twitter.com/MuzitoAdolphe	N/A

A screengrab showing the data schema for mapping DRC presidential candidates' Facebook, Instagram, TikTok, and X profiles (Source: CfA using PIPs data)

id	name_id	full_name	tiktok_url	note_yt	YouTube_url	note_web	website	note_wiki	wikipedia_url	wa_numbers	telegram
pers01	radjabho_tebabho__soborabo	Radjabho Tebabho Soborabo	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers02	mutamba_tungunga_constant	Mutamba Tungunga Constant	https://www.tiktok.com/@mutar	N/A	N/A	official	http://www.nogec.org/	N/A	N/A	N/A	N/A
pers03	katumbi_chapwe_moise	Katumbi Chapwe Moise	N/A	campaign	https://www.youtube.com/@mo	official	https://ensemble-rdc.com/	wikipedia	https://en.wikipedi	N/A	N/A
pers04	sesanga_hipungu_dja_delly	Sesanga Hipungu Dja Delly	https://www.tiktok.com/@dsesa	N/A	N/A	official	https://envoirdcongo.org/	N/A	N/A	https://api.whatsapp.com/send?	N/A
pers05	anzuluni_isiloketshi_floribert	Anzuluni Isiloketshi Floribert	N/A	N/A	N/A	official	https://nyfalm-a.com/	N/A	N/A	N/A	N/A
pers06	baende_etafe_eliko_jean_claude	Baende Etafe Eliko Jean Claude	N/A	N/A	N/A	official	https://www.jeanclaudebae	N/A	N/A	N/A	N/A
pers07	bolamba_tony_cassius	Bolamba Tony Cassius	N/A	N/A	N/A	N/A	N/A	N/A	N/A	https://api.whatsapp.com/send?	N/A
pers08	buse_falay_georges	Buse Falay Georges	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers09	ikofu_mputa_mpunga_marie_josse	Ikofu Mputa Mpunga Marie-Josse	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers10	kazadi_kanda_rex	Kazadi Kanda Rex	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers11	kikuni_masudi_seth	Kikuni Masudi Seth	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers12	majondo_mwamba_patrice	Majondo Mwamba Patrice	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers13	masalu_anedu_andre	Masalu Anedu Andre	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers14	mudekereza_bisimwa_justin	Mudekereza Bisimwa Justin	N/A	N/A	N/A	campaign	https://vraichangement.org/	N/A	N/A	N/A	N/A
pers15	mukwege_mukengere_denis	Mukwege Mukengere Denis	N/A	N/A	N/A	official	https://panzifoundation.org	wikipedia	https://en.wikipedi	N/A	N/A
pers16	ngalasi_kurisini_aggrey	Ngalasi Kurisini Aggrey	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers17	ngoy_ilunga_wa_theodore	Ngoy Ilunga Wa Theodore	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pers18	nkema_liloo_bokonzi_loli	Nkema Liloo Bokonzi Loli	N/A	N/A	N/A	official	http://lolinkema.com/	N/A	N/A	N/A	N/A
pers19	tshiani_k_musadiamvita_noel	Tshiani K Musadiamvita Noel	N/A	N/A	N/A	N/A	N/A	N/A	N/A	+243852102222	N/A
pers20	tshisekedi_tshilombo_felix_antoine	Tshisekedi Tshilombo Felix Antoine	N/A	N/A	N/A	N/A	N/A	wikipedia	https://en.wikipedi	N/A	N/A
pers21	fayalu_madidi_martin	Fayalu Madidi Martin	N/A	N/A	N/A	N/A	N/A	wikipedia	https://en.wikipedi	N/A	N/A
pers22	matata_ponyo_mapon	Matata Ponyo Mapon	https://www.tiktok.com/@matat	campaign	https://www.youtube.com/@ma	campaign	https://matataponyomapon	wikipedia	https://en.wikipedi	+243812763003	N/A
pers23	diongo_shamba_franck	Diongo Shamba Franck	https://www.tiktok.com/@mlprc	campaign	https://www.youtube.com/@mlp	campaign	https://mlprdc.org/	N/A	N/A	+243 900 234 043	N/A
pers24	muzito_fumutshi_adolphe	Muzito Fumutshi Adolphe	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

A screengrab showing the data schema for mapping DRC presidential candidates' website, Wikipedia, WhatsApp, Telegram and YouTube profiles (Source: CfA using PIPs data)

The data schema includes columns for the candidates' unique identifier (ID), full name, gender, political party, party ID, party abbreviation, and the office sought. Accurate naming is crucial to ensure correct social media mapping.

Once verified, profiles on platforms such as Facebook, Instagram, Telegram, TikTok, X, and YouTube are mapped. Accounts are classified as campaign or personal, with a focus on campaign accounts, as they primarily drive election agendas.

Among PIPs and entities, each group plays a role in shaping the political landscape. Understanding their influence and interactions is essential to assessing their impact on election outcomes.

A screengrab showing the classification of DRC presidential candidates' social media accounts as either personal or campaign (Source: CfA using PIPs data)

Categorisation of PIPs

This is a crucial step in social media monitoring, particularly when creating a watchlist to track conversations and set up early warning systems. Here is a brief overview of how PIPs can be categorised:

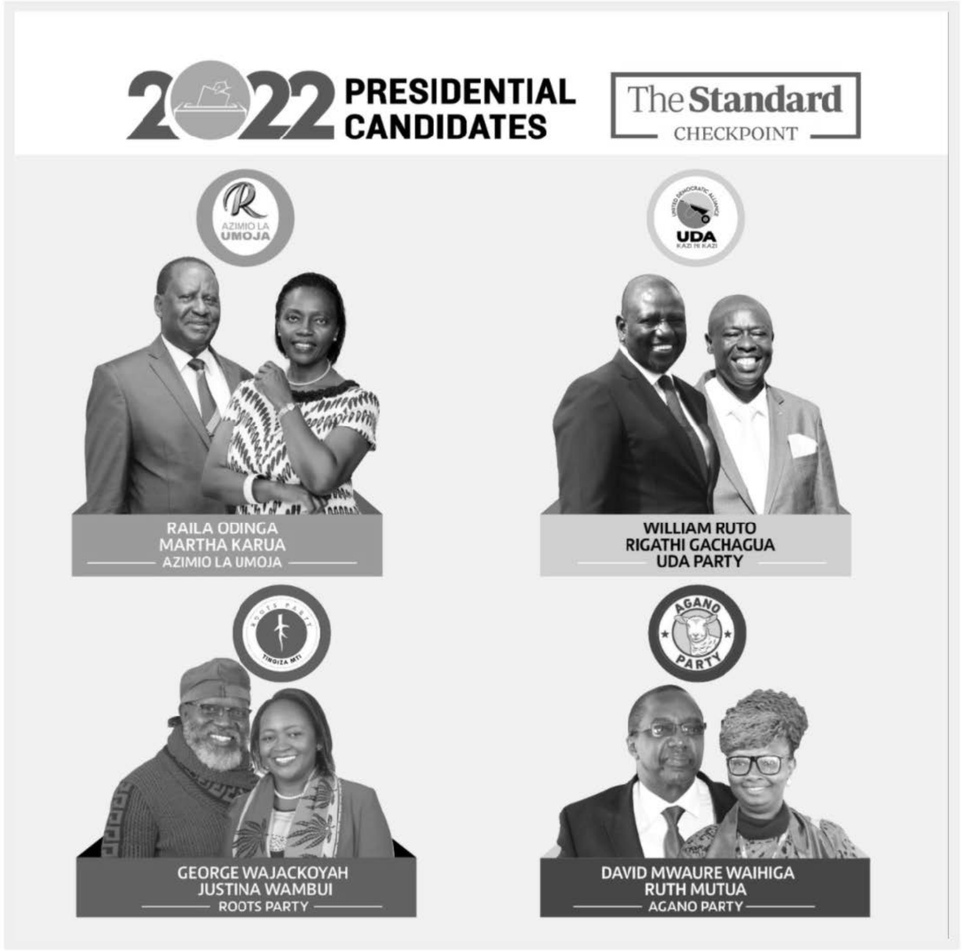
Presidential candidates

Presidential candidates are PIPs, representing their parties and leading major campaign efforts. Their support networks, consisting of advisers, party leaders, and strategists, play crucial roles in shaping campaign narratives. When mapping candidates, it is essential to include their running mates (where applicable), as they can influence voter support by representing strategic demographic, geographic, or political interests.

Independent candidates, lacking party infrastructure, rely on their personal brands and direct voter engagement, making their mapping distinct yet equally vital.

Spouses of presidential candidates, if politically active or publicly influential, play a role in advocating for campaign policies, rallying voter bases, and shaping perception. Their involvement can impact campaign dynamics and voter sentiment, making them essential to map alongside candidates.

Election periods generate vast data on PIPs from official websites, social media, and Wikipedia, revealing campaign strategies, narratives, and shifts in public discourse. Monitoring these sources helps track candidates' messaging, political dynamics, and voter engagement.



Screenshot showing the list of presidential candidates and their running mates during the 2022 Kenyan general elections (Source: The Standard Newspaper)

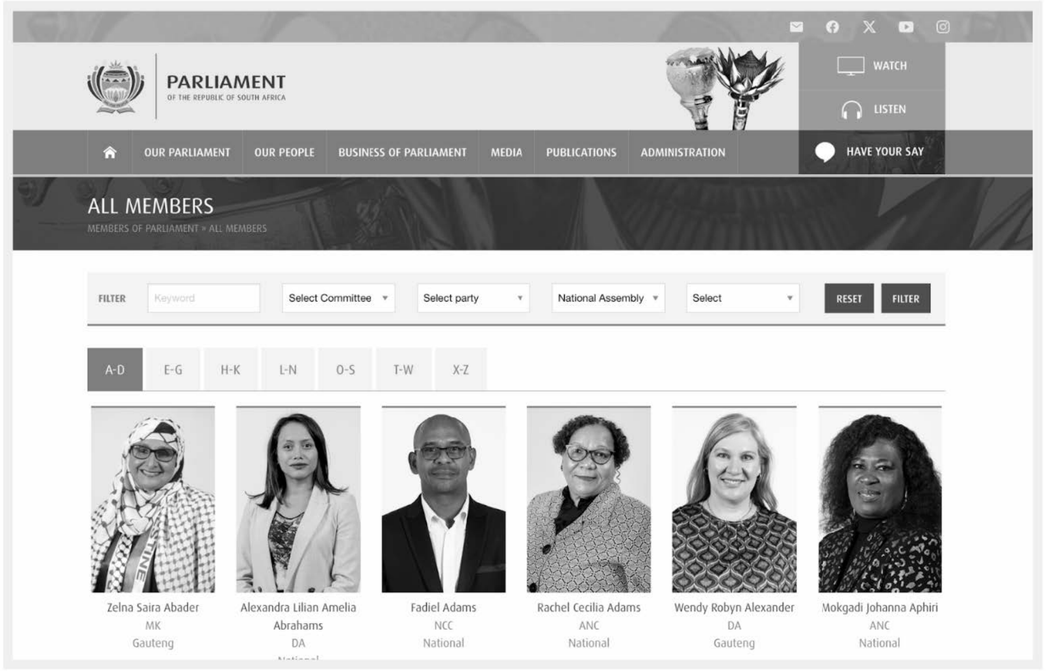
Incumbents

Incumbents such as – sitting governors, legislators, and presidents – are essential during elections, as many may run for office. Early identification enables close monitoring of their campaign strategies, messaging, and voter engagement, particularly through social media. Electoral body websites, parliamentary databases, and public records offer valuable data for tracking incumbents' activities.

Tracking incumbents is crucial to understanding their influence in both digital and physical political spaces. Those seeking re-election often leverage their positions to shape the political landscape through:

i. **Creation of 'zoning':** Blocking rival candidates from campaigning in some regions to maintain dominance.
ii. **Politically charged violence:** Employing intimidation or orchestrating unrest to suppress opposition and voter turnout.
iii. **Spearheading smear campaigns:** Coordinating the spread of false, misleading, or negative information about opponents to discredit them.

Online discourse often foreshadows offline electoral misconduct, from ballot buying to voter suppression. Monitoring social media provides early warning signs of emerging tensions, helping to anticipate and mitigate risks before they escalate.



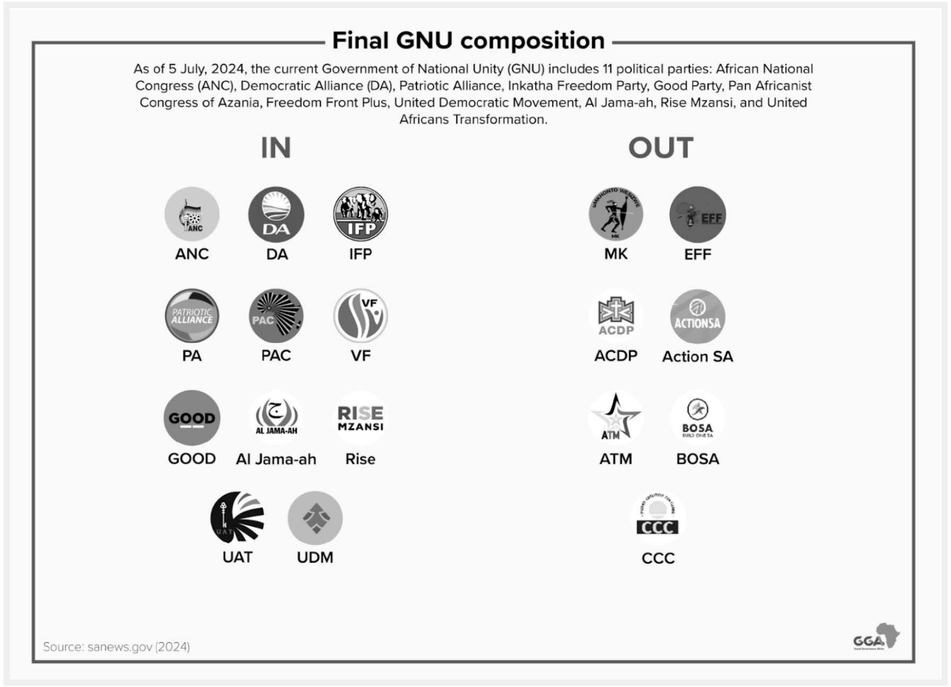
A screenshot showing the incumbent legislators on the South African parliamentary website (Source: CfA via the SA parliament)

Opposition figures

Opposition politicians and officials shape the electoral landscape by mobilising dissent and challenging incumbents. Like those in power, they engage in behaviours that influence public perception:

i. **Smear campaigns:** Opposition figures often target to discredit incumbents while promoting their own ideologies, highlighting government failures to gain support.
ii. **Dissemination of misleading facts:** The figures may spread half-truths or misleading narratives to undermine public confidence in the governing or other rival parties, contributing to electoral disinformation.

Tracking their online activity on blogs, internet forums, or social media platforms reveals emerging opposition voices and the traction of their narratives. In parliamentary systems with official opposition roles, parliamentary websites provide additional insights into figures, enriching political analysis.



A screenshot providing an overview of the current opposition parties in South Africa after the May 2024 elections (Source: CfA using Good Governance)

Military officials

In countries where the military dominates politics, high-ranking officers often wield more power than elected officials, especially in countries with histories of coups, such as Burkina Faso, Mali, and Niger. Their control over security forces and state institutions allows them to shape governance, often operating behind the scenes to avoid scrutiny.

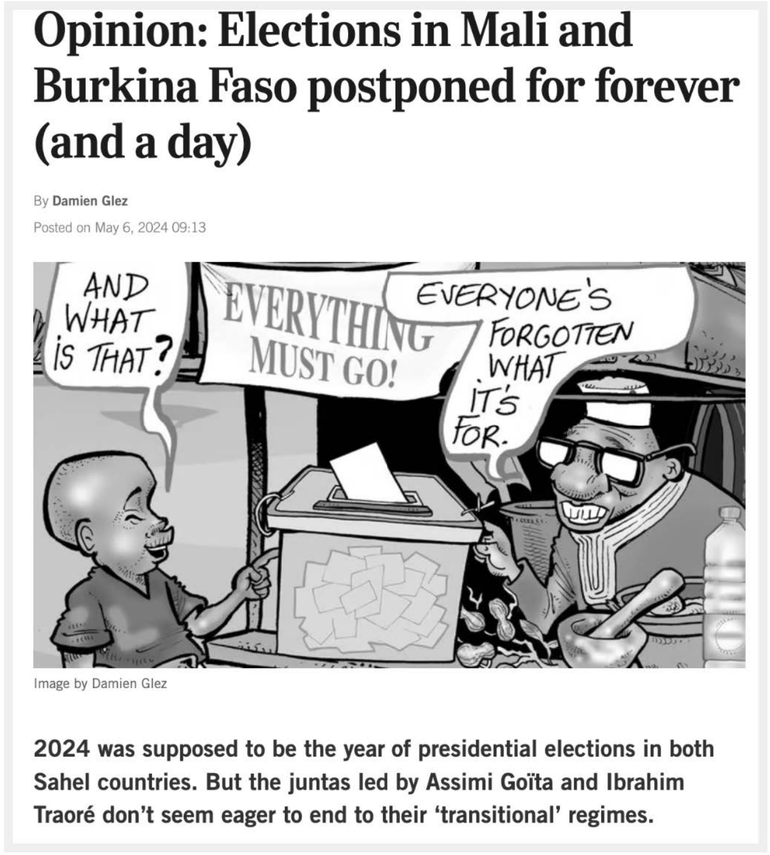
One challenge in tracking military-backed governments is the scarcity of publicly available data on top generals. Unlike politicians, they lack public profiles, making it difficult to map their influence.

Election periods in such regimes are often marked by intimidation and suppression, discouraging opposition and ensuring limited competition. The controlled political environment consolidates military power.

Strategies to track military influence include:

i. **Monitoring official statements:** Press conferences, social media updates, and state media broadcasts offer clues about military officials' agendas.
ii. **Following international engagements:** Military leaders may be more visible in regional or global meetings, revealing alliances and political positioning.

Though personal details remain elusive, analysing these appearances helps map the role of military officials in governance.



Opinion: Elections in Mali and Burkina Faso postponed for forever (and a day)

By **Damien Glez**

Posted on May 6, 2024 09:13

Image by Damien Glez

2024 was supposed to be the year of presidential elections in both Sahel countries. But the juntas led by Assimi Goïta and Ibrahim Traoré don't seem eager to end to their 'transitional' regimes.

A screenshot highlighting the challenges junta-led countries face in holding elections to return to civilian rule (Source: CfA via The Africa Report)

Labour union leaders

Labour union leaders wield significant political influence, often aligning with parties and mobilising workers at scale. It is critical to identify vocal union leaders shaping political discourse, and track unions with large memberships influencing voter behaviour. While their power is largely offline – via negotiations, rallies, and strikes – many use social media to extend their reach. Union websites and social media accounts provide insights into their political activities.

WORKING TOGETHER

Kenya: Can workers unions and COTU help Raila win presidency in 2022?

By **Victor Abuso**

ar Reserved for subscribers

Posted on January 5, 2022 09:11

A supporter of Kenya's Opposition leader Raila Odinga walks past his election posters before the Azimio la Umoja (Declaration of Unity) rally to unveil his August 2022 Presidential race candidature at the Moi International Sports centre in Kasarani, Nairobi, Kenya 10 December 2021. REUTERS/Baz Ratner

The leadership of Kenya's Central Organisation of Trade Unions (COTU), a national umbrella body of trade unions, says it is endorsing veteran opposition leader Raila Odinga's bid to vie for presidency on 9 August this year.

An example from Kenya's 2022 general elections, where labour union leaders were at the centre of presidential politics (Source: CfA via The Africa Report)

3.1.4. How to build lexicons and narratives

Listening and monitoring systems

Monitoring or listening systems are methods used to track and find information online. Monitoring systems can be as simple as performing a keyword search using the right phrases on different social platforms.

Types of monitoring and listening systems

Some of the listening and monitoring systems one can set up to track information online include:

i. Creating X Facebook and X lists.
ii. Following multiple social and political Facebook groups, especially those with large followings.
iii. Monitoring the trending X topics.
iv. Setting up Google Alerts.
v. Joining several WhatsApp groups.
vi. Crowdsourcing, which involves asking people to share any suspicious information they see online through an email, tipline, web online form or social media platforms.

Importance of monitoring/listening systems

i. **Capture unseen posts:** Even with expertise in social media investigations, it is impossible to spot every instance of misinformation within your region. Monitoring systems, such as Google Alerts, help you capture posts you might otherwise miss by sending email alerts when keywords such as 'beware', 'fake', or 'scam' are used.

ii. **Diversification:** Focusing on a specific area may limit you to a particular trend, such as hoaxes. Monitoring systems allow for broader coverage, enabling you to track misinformation on various topics. For instance, using a Google form or WhatsApp tipline, you can ask your audience to report potential misinformation, resulting in a diverse range of claims for debunking. Here is how to create a Google form.

iii. **Filtering information:** With millions of daily social media posts, manual review is impractical. Monitoring systems filter data to match your search criteria. Features such as advanced X search allow you to filter posts by account, language, and timeframe, ensuring a focused and relevant dataset.

iv. **Saves time:** Monitoring systems streamline the process of identifying claims. Using Facebook or X lists, you can group accounts known for spreading misinformation and quickly check their posts without scrolling through individual timelines. Steps to create Facebook and X lists are outlined here and here.

Selecting and implementing monitoring tools

Choosing the right monitoring tool is crucial for detecting mis-/disinformation.
Below are two criteria for selecting and implementing monitoring tools.

i. **Examine tool functionality:** When choosing monitoring tools, assess their functionalities to determine which best meets your needs. Different tools serve different purposes – while a Facebook list helps track specific accounts, Google Alerts notifies you when monitored keywords appear online. A fact-checking organisation might use a Facebook list to track misinformation sources, whereas a public relations firm may rely on Google Alerts for real-time updates on their brand or clients.

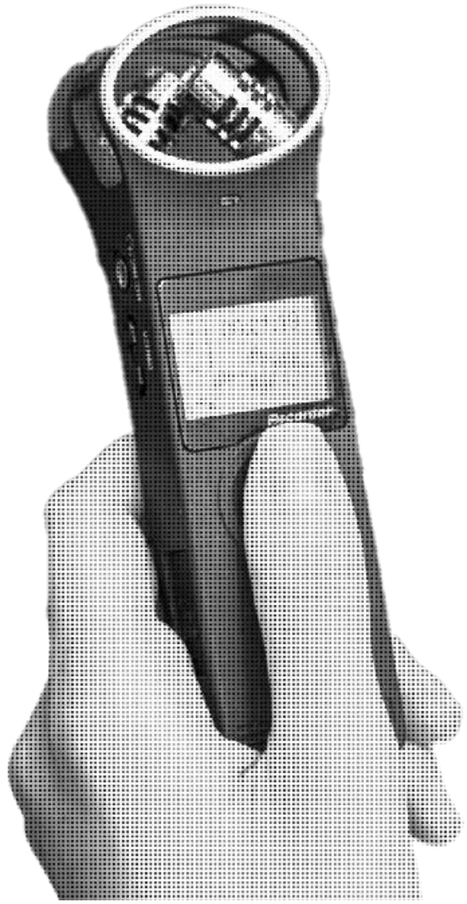
ii. **Cost implication:** Subscribing to monitoring tools can be quite costly. Therefore, it is advisable that you use open-source tools as they are free and easily accessible. Facebook lists, Google Alerts, and X lists are free.

Ethical considerations when selecting monitoring tools

When creating a monitoring system, it is crucial to address the following ethical concerns of data protection, privacy, and transparency:

i. **Transparency:** Clearly define the purpose of your monitoring system and outline the methodologies and expected outcomes. This ensures stakeholders understand the intent behind your efforts and prevents the misinterpretation of monitoring as an invasion of rights.

ii. **Privacy and data protection:** Respect individuals' right to privacy. For example, refrain from fact-checking content not shared publicly. Additionally, ensure that information from sources, such as emails, Google Forms, or tips from a tipline, remains anonymous to protect whistleblowers from potential retaliation.



Monitoring lexicons

Lexicons are structured databases of keywords, phrases, and thematic terms used to track narratives, trends, and entities in online spaces. They play a crucial role in monitoring disinformation, electoral discourse, and hate speech by identifying 'trigger' terms that incite audiences.

Creating a lexicon involves selecting highly relevant terms aligned with research objectives, whether tracking coordinated campaigns, elections, or specific hashtags. A well-compiled lexicon enhances targeted analysis, ensuring precise interpretation of online conversations.

Lexicons evolve with digital discourse, adapting to emerging terminologies and patterns to maintain their effectiveness in media monitoring and investigative work.

The table to the right outlines the elements of lexicons:

Classifier	Description
Keywords	These are context-specific words or phrases that carry particular significance in digital conversations.
Terms	Relevant keywords or phrases that actors often weaponise to attack, demean, or insult groups or individuals. These include derogatory terms, slurs, and coded language that dehumanises, incites hostility, or reinforces stereotypes. Their impact varies based on cultural and contextual factors but is consistently aimed at escalating conflict and division.
Variant terms	Synonyms or variations of terms that carry the same meaning but are used in different contexts, dialects, or pronunciations.
Language	Refers to the medium of communication, encompassing spoken, written, and digital expressions in languages such as Arabic, English, Igbo, Kiswahili, and Shona.
Offensiveness	A severity scale assigned to words and phrases based on their level of unpleasantness, ranging from mild insults to highly aggressive slurs. The impact of such language varies depending on cultural context, intent, and the targeted group, helping to assess the degree of harm or inflammatory potential in communication.
Regions	Refers to geographical areas associated with specific ethnic languages or communities. These locations act as cultural and linguistic hubs, shaping contexts, dialects, social norms, and traditions.
Content	Refers to the circumstances surrounding a message, shaping its meaning and impact. Factors such as ancestry, class, ethnicity, extremism, gender, nationality, race, religion, slurs, violence, and xenophobia influence how a message is perceived. These elements determine tone, intent, and the potential for escalation or harm in communication.

The data schema for classifying lexicons provides a structured approach to organising and analysing language patterns. It includes factors to consider when building lexicons including:

i. **Context**

At its core, the schema incorporates an index column, which assigns a unique identifier to each entry, allowing for systematic tracking and reference.

Additionally, a category column serves as the foundation for classification, helping contextualise terms based on their relevance to different socio political themes.

Index	Category	Description	Note	
cat_01	Class	Keywords based on classes in the society.		
cat_02	Ethnicity/ Ethnic Slur	Keywords that are used to target people of other ethnic background or trib		
cat_03	Nationality	Keywords that are used to target people of other nationalities with an aim		
cat_04	Race	Keywords that are used to target people of other races with an aim to pror		
cat_05	Stereotype	Keywords that are used to target people based on their physical appearar		
cat_06	Ancestry	Keywords based on ancestry such as kingdoms, which when used they pr		
cat_07	Gender	keywords that are targeted towards a particular gender with an aim to und		
cat_08	Xenophobia	This includes keywords that show prejudice against people from other cou	Tutsis in Both Rwanda and DRC	
cat_09	Extremism	keywords that promote extremism.		
cat_10	Violence	Keywords whenever used are meant to stir up violence tension.		
cat_11	Deragatory	Keywords that are disrespectful and hostile criticism.		

A screengrab of an example of a context category to guide in lexicon identification
(Source: CfA using DRC lexicon data)

These categories are not exhaustive and may shift depending on geographical and cultural contexts.

Classifications include class, ethnicity, ethnic slurs, and religion, which capture language reflecting racial discrimination, religious bias, or social hierarchies. Colour, nationality, and race shape discourse on identity, inclusion, and exclusion, while ancestry can highlight heritage or be weaponised for exclusion.

Gender-based language reveals societal attitudes, including gendered insults, misogyny, and patriarchy. Extremism and xenophobia propagate hate, ideological bias, or radicalisation, while violence-related terms highlight rhetoric inciting harm or conflict.

ii. Region

Geographical classification of lexicons helps capture regional language nuances, reducing false positives and improving the identification of harmful content. Words may have different meanings across regions, making context crucial in distinguishing between neutral and harmful usage. For example, in the DRC, lexicons were first categorised by province and then refined by region. Assigning an index number based on geography allowed for a more precise classification of ethnic terms, ensuring accurate contextual analysis.

Index	Provinces	Region	Note
Prov_001	Kasai Province	Central Region	prone to conflict region
Prov_002	Kasai-Central Province	Central Region	prone to conflict region
Prov_003	Kasai-Oriental Province	Central Region	
Prov_004	Sankuru Province	Central Region	
Prov_005	Ituri Province	Eastern Region	prone to conflict region
Prov_006	Haut-Uele Province	Eastern Region	prone to conflict region
Prov_007	Tshopo Province	Eastern Region	prone to conflict region
Prov_008	Bas-Uele Province	Eastern Region	prone to conflict region
Prov_009	Haut-Lomami Province	Eastern Region	prone to conflict region
Prov_010	North Kivu Province	Eastern Region	prone to conflict region
Prov_011	South Kivu Province	Eastern Region	prone to conflict region
Prov_012	Maniema Province	Eastern Region	prone to conflict region
Prov_013	Tanganyika Province	Eastern Region	prone to conflict region
Prov_014	Kinshasa Province (the capital c	Western Region	

A screengrab of an example showcasing the mapping of DRC regions and provinces (Source: CfA using the DRC lexicon database)

iii. Key term classification

Key term classification is essential when building a lexicon database. Terms are categorised as phrases or words to refine monitoring and query development. Offensive classification assigns a scale to gauge derogatory or harmful usage. Language and location categorisation ensures recognition of regional variations. Context is crucial in assessing offensiveness, especially during elections on platforms such as Facebook, Instagram, TikTok, and X.

Example of key term classification in the Kenyan context:

- **Keyword:** Madoadoa (Swahili word for 'blemishes').
- **Classification:** Highly offensive, used nationally by some politicians.
- **Context:** Used to incite violence or promote ethnic cleansing by targeting non-inhabitants of a specific region or political jurisdiction. In political discourse, it polarises communities.
- **Classification in Kenya:** Considered a hate lexicon due to its potential to incite violence and promote division.

Madoadoa		Word	Highly offensive
Madude		Word	
Mageryenge		Word	
Mende		Word	
Mombasa ni yetu wabaara wa		Phrase	
Muhajir		Word	

Extrememly offensive
Highly offensive
Mildly offensive
Moderately offensive

A screenshot of the classification of the hate term 'madoadoa' using a scale to show its level of offensiveness (Source: CfA using the Kenya lexicon database)

Lexicon building based on identity factors

Lexicon building involves systematically compiling terms and phrases that reflect identity categories such as ethnicity, nationality, race, region, and sex. These distinctions help capture the nuances of language use across different social and cultural settings, shaping communication patterns, social perceptions, and group identities.

Term	Variant	Type	Offensiveness	Language	location	I2 ntext	Des
Kimurkeldet	Kimurkelda	Word	Highly offensive	Kalenjin	Rift valley	Ethnicity/ Ethnic Slur	Thi
Kura haramu, kura ukafiri		Phrase	Highly offensive	Swahili	Coastal	Class	
Kwekwe	makwekwe	Word	Highly offensive	Swahili	Coastal	Religion	
Lazima tushinde aidha kupitia		Phrase	Mildly offensive	Swahili	Kenya	Ethnicity/ Ethnic Slur	
Madoadoa		Word	Highly offensive	Swahili	Kenya	Nationality	
Madude		Word	Mildly offensive	Swahili	Coastal	Race	
Mageryenge		Word	Mildly offensive	Mijikenda	Coastal	Colour	
Mende		Word	Mildly offensive	Swahili	Coastal	Ancestry	
Mombasa ni yetu wabaara wa		Phrase	Mildly offensive	Swahili	Coastal	Gender	
Muhajir		Word	Mildly offensive	Kenyan Arabs	Coastal	Xenophobia	
Munafikin	munafik	Word	Mildly offensive	Swahili	Coastal	Extremism	
Muslims are al-Shabaab		Phrase	Highly offensive	English	Kenya	Violence	
Ngawira		Word	Mildly offensive	Swahili	Coastal		
Nitakutoa supu (kutoa damu); Ntakutoa Supu, kunywa dami ya mtu);	Hatuwawachi hadi tumwage damu	Phrase	Extrememly offensive	Swahili	Kenya		

An overview of the lexicon schema with the various classifications. (Source: CfA using the Kenya lexicon database)

Below is a breakdown of key terms related to identity and classification, including ethnicity, nationality, race, region, and sex. These terms help to understand the various ways people are categorised and how these categories intersect in different social, political, and cultural contexts.

i. **Gender:** Words and phrases related to sex encompass gender-specific language, pronouns, and terms reflecting identities, social roles, or stereotypes. A lexicon in this category helps identify discrimination or misogyny, monitor gender portrayals, and track gendered language patterns.

ii. **Ethnicity:** Ethnic-specific lexicons encompass terms tied to ethnic groups, identities, and traditions. They help track culturally significant phrases, divisive language, and ethnic slurs. For instance, negative stereotypes linking communities to traits such as laziness or witchcraft reinforce harmful biases.

iii. **Regional:** Regional lexicons track terms unique to specific areas, including slang and dialects. They help identify location-based narratives and variations in meaning across regions. For example, phrases common in West Africa may have different connotations in East Africa, influencing interpretation.

iv **Nationality:** Nationality-based lexicons capture terms linked to national identity, stereotypes, and xenophobia. They help track references to countries, national figures, and symbols, especially when used to incite hostility or reinforce bias.

v. **Race:** Race-related lexicons identify terms linked to racial identity, racialised language, and slurs. They are essential for analysing racial discourse, monitoring hate speech, and tracking racist rhetoric.

Lexicon building for event-based terms

Event-based lexicons are specialised collections of phrases, terminologies, and words tied to specific events, occurrences, or time-bound situations. These terms capture the language and narrative patterns surrounding events, making them essential for social media monitoring and analysis. They help track and interpret how public discourse shapes responses to elections, major incidents, protests, social movements, and other impactful occurrences.

Considerations for event-based lexicon building:

i. **Identifying event-specific keywords:** Terms linked to particular events, such as the names of political gatherings, protest slogans, or trending hashtags, are central to this type of lexicon. These keywords help monitor how events are framed, perceived, and reported in real time.

ii. **Relevance:** Some phrases are only relevant during the event itself, while others evolve in meaning as the situation unfolds. Event-based lexicons must be continuously updated to reflect these changes. For example, the hashtag #EndSARS gained significance during the Nigerian anti-government protests that took place between 01 and10 August 2024, but later took on broader connotations in discussions on police brutality.

iii. **Emotion and sentiment:** Events often generate emotionally charged language. Lexicons should include terms that capture a range of sentiments, such as outrage or calls to action (e.g., burn houses), as these significantly influence public opinion and narrative framing.

iv **Monitoring the evolution of events:** The language surrounding an event can shift rapidly. A peaceful protest may escalate into violent clashes, changing the terms used to describe it. An effective event-based lexicon must be adaptable to these shifts. For instance, In Kenya, the #RejectFinanceBill2024 movement began as a peaceful anti-government outcry on social media in May 2024. However, by 18 June 2024, the protests, primarily led by Generation Z, escalated into nationwide demonstrations that turned violent due to confrontations with police over a month-long period. As the protests gained momentum, counter-narratives emerged, shifting the discourse to anti-Gen Z sentiment. Opponents sought to discredit the movement by associating Gen Z protesters with the LGBTQ+ community, using derogatory hashtags such as #GayZ to frame the demonstrators as being members of the LGBTQ+ community.

Hate lexicon

A hate lexicon is a curated collection of coded language, keywords, phrases, and slurs commonly used in hate speech. It serves as a critical tool for identifying and categorising harmful language in both online and offline spaces.

The United Nations (UN) defines hate speech as any form of communication – behavioural, spoken, or written – that discriminates against or attacks individuals or groups based on identity factors such as ethnicity, colour, descent, gender, nationality, race, or religion. There is no universal definition of hate speech under international human rights law, as it remains contested in relation to equality, freedom of expression, and non-discrimination.

To build an effective hate lexicon, leveraging expertise from specialised organisations is essential. These organisations focus on identifying hateful words, slang, and coded language across different regions, making them valuable resources in lexicon development. Some initiatives include: Dangerous Speech Project, Hatebase, Masakhane Project and PeaceTech Lab.

Components of a hate lexicon

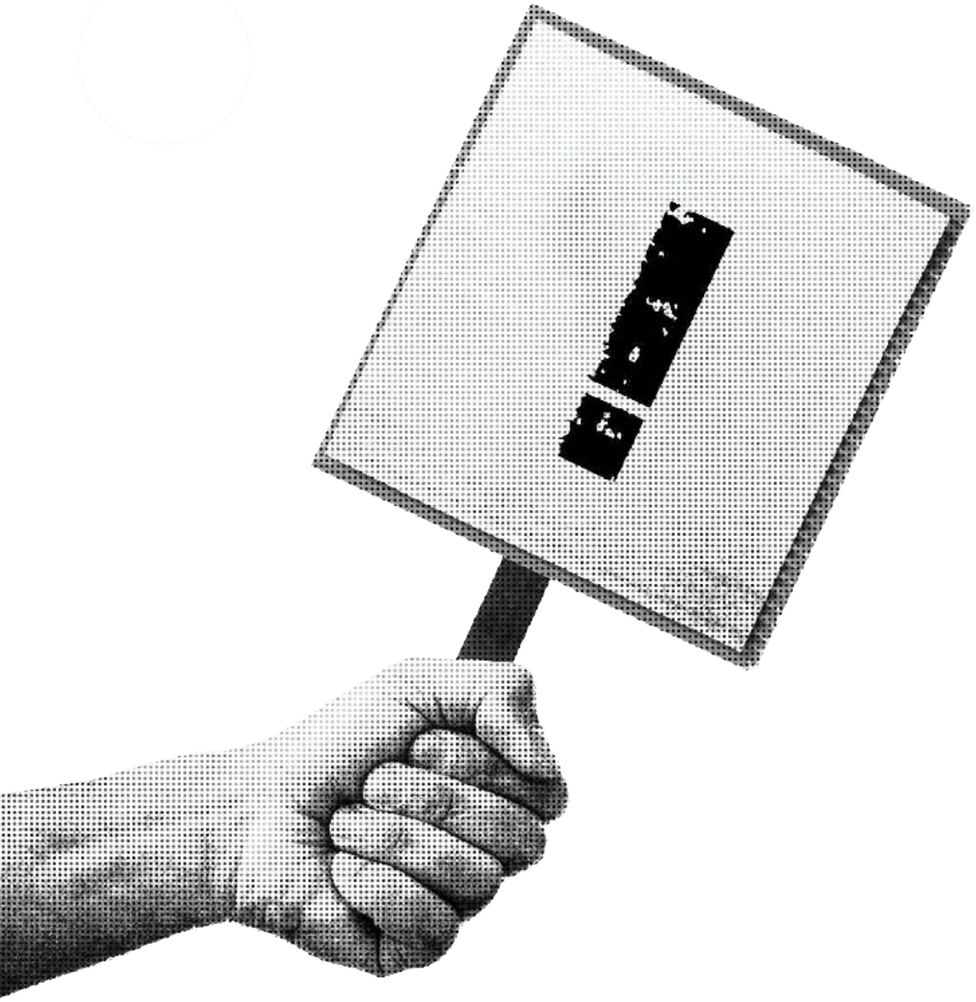
i. **Hate speech terms:** These are explicit words or slurs used to attack, demean, or insult individuals or communities. They may include anti-LGBTQ+ phrases that are offensive, misogynistic terms, or racial slurs.

	Index	Term	Variant	Type	Offensiveness	Language	location	Context
1	Ter_01	Kimurkeldet	Kimurkelda	Word	Highly offensive	Kalenjin	Rift valley	Ethnicity/ Ethnic S
2	Ter_02	Kura haramu, kura ukafiri		Phrase	Highly offensive	Swahili	Coastal	Religion
3	Ter_03	Kwekwe	makwekwe	Word	Highly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
4	Ter_04	Lazima tushinde aidha kupitia		Phrase	Mildly offensive	Swahili	Kenya	Extremism
5	Ter_05	Madoadoa		Word	Mildly offensive	Swahili	Kenya	Ethnicity/ Ethnic S
6	Ter_06	Madude		Word	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
7	Ter_07	Mageryenge		Word	Mildly offensive	Mijikenda	Coastal	Ethnicity/ Ethnic S
8	Ter_08	Mende		Word	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
9	Ter_09	Mombasa ni yetu wabaara wa		Phrase	Mildly offensive	Swahili	Coastal	Ethnicity/ Ethnic S
10	Ter_10	Muhajir		Word	Mildly offensive	Kenyan Arabs	Coastal	Religion
11	Ter_11	Munafikin	munafik	Word	Mildly offensive	Swahili	Coastal	Religion
12	Ter_12	Muslims are al-Shabaab		Phrase	Highly offensive	English	Kenya	Religion

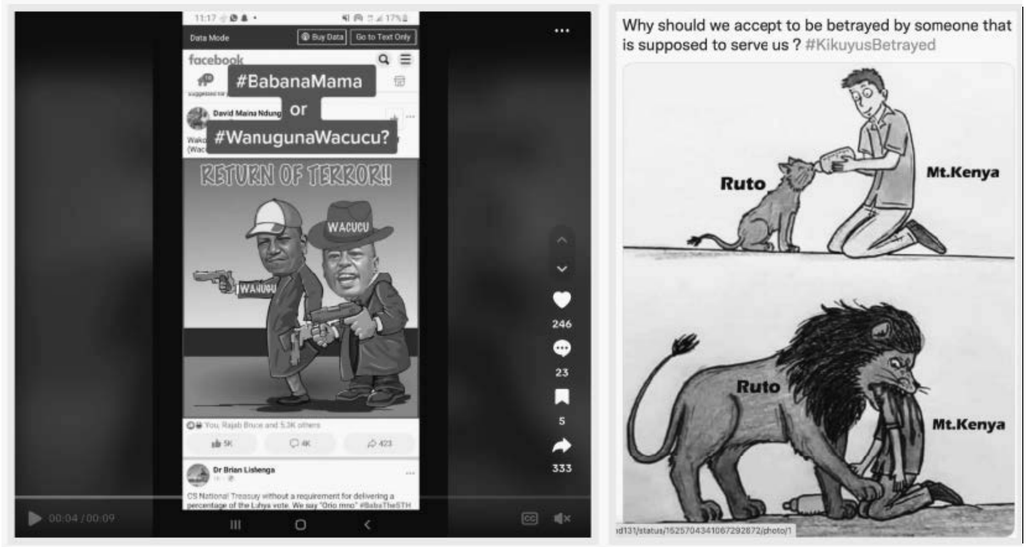
A screenshot showing a collection of Kenyan hate terms with various classifications (Source: CfA using the Kenya lexicon database)

ii **Coded language:** Malign actors often use coded language to avoid detection. These can include seemingly innocent words with hidden meanings or terms that have been reappropriated to convey hate within specific communities.

A screenshot highlighting the use of coded language during the Nigeria elections (Source: Facebook)



iii. **Use of imagery:** Hate speech is not always limited to words. It can also include specific phrases, slogans, or even symbols that convey hostile intent. Capturing these expressions and visual elements (e.g., hate symbols or imagery) expands the lexicon's capability to detect forms of hateful content.



A screenshot showing the use of imagery to spread hateful content during the 2022 Kenyan elections (Source: CfA via ADDO)

A dynamic hate lexicon must be continuously updated to track these shifts, ensuring real-time identification and mitigation of harmful content.

Freedom of expression vs hate speech

The International Covenant on Civil and Political Rights defines freedom of expression as the right to hold opinions without interference, including the right to seek, receive, and share information across borders through any media.

The Rabat Plan of Action sets a high threshold for restricting freedom of expression, particularly regarding incitement to hatred. It applies a six-part test to assess whether speech warrants limitation: context, content and form, extent of dissemination, intent, likelihood and imminence of harm, and speaker.

(1) **Context**: Context is of great importance when assessing whether particular statements are likely to incite discrimination, hostility or violence against the target group, and it may have a direct bearing on both intent and/or causation. Analysis of the context should place the speech act within the social and political context prevalent at the time the speech was made and disseminated;

(2) **Speaker**: The speaker's position or status in the society should be considered, specifically the individual's or organization's standing in the context of the audience to whom the speech is directed;

(3) **Intent**: Article 20 of the ICCPR anticipates intent. Negligence and recklessness are not sufficient for an act to be an offence under article 20 of the ICCPR, as this article provides for "advocacy" and "incitement" rather than the mere distribution or circulation of material. In this regard, it requires the activation of a triangular relationship between the object and subject of the speech act as well as the audience;

(4) **Content and form**: The content of the speech constitutes one of the key foci of the court's deliberations and is a critical element of incitement. Content analysis may include the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed in the speech or the balance struck between arguments deployed;

(5) **Extent of the speech act**: Extent includes such elements as the reach of the speech act, its public nature, its magnitude and size of its audience. Other elements to consider include whether the speech is public, what means of dissemination are used, for example by a single leaflet or broadcast in the mainstream media or via the Internet, the frequency, the quantity and the extent of the communications, whether the audience had the means to act on the incitement, whether the statement (or work) is circulated in a restricted environment or widely accessible to the general public; and

(6) **Likelihood, including imminence**: Incitement, by definition, is an inchoate crime. The action advocated through incitement speech does not have to be committed for said speech to amount to a crime. Nevertheless, some degree of risk of harm must be identified. It means that the courts will have to determine that there was a reasonable probability that the speech would succeed in inciting actual action against the target group, recognizing that such causation should be rather direct.

The Rabat Plan of Action's six-part threshold test for restricting freedom of expression in detail (Source: the Rabat Plan of Action)

3.2. Some frameworks for analysing IMI

Information manipulation and interference (IMI) campaigns can often be complex, making them difficult to understand, analyse, and explain. Because of this, some frameworks have been created to ease the process of detecting and labelling such campaigns by paying attention to their component parts. They also make it easier to coordinate responses to the IMI campaigns.

3.2.1. Listening and monitoring systems

The ABCDE Framework is a system used to detect and label information manipulation and interference (IMI) by identifying five elements: `

• Actors • Behaviour • Content • Degree • Effect.

Actor

These are the central figures in an IMI incident, including proxies operating within or beyond their territory. Identifying actors requires digital forensics, investigative methods, and robust attribution frameworks to establish responsibility and intent.

The table below highlights the questions to ask in identifying actors in an IMI incident:

Actor type	Question
Individual(s)	Is the person involved acting in their private capacity?
Non-state actor(s)	Is the actor affiliated with a private or an organisation not linked to a government?
Media platform(s)	To what degree is the platform of distribution independent?
Political actor(s)	Does the individual act on behalf of a recognised political entity?
Foreign state(s)	Is the actor an agent or proxy of a foreign government?

Behaviour

Behaviour refers to the strategy used to execute an IMI incident. While some behaviours are straightforward, threat actors often employ complex techniques to evade detection. Identifying deceptive behaviour at scale requires data analysis and amplification pattern studies. Access to reliable data is essential for effective detection and response.

The table below outlines the questions to ask in identifying behaviour in an IMI incident:

Behaviour type	Question
Transparency	Is the actor disguising their identity or actions?
Authenticity	Is the actor using illegitimate communication techniques?
Infrastructure	Is there evidence of behind-the-scenes coordination?
Intent	Does the behaviour suggest a malign intent?

Real-life example:
Uganda's 2021 general elections

In the lead-up to Uganda's 2021 elections, Facebook dismantled several fake accounts linked to the Ugandan government, which were being used to distort online discourse and mislead the public. These accounts impersonated real users, engaged with posts to fabricate the appearance of widespread support for the ruling party, and amplified government narratives. Beyond social media manipulation, digital authoritarian tactics such as website blockages and restrictions on SMS services were also employed to control information flow. By leveraging state-controlled media, these efforts sought to shape voter perceptions and suppress dissenting voices.

Click here to read the BBC article.

Content

Content refers to the message or narrative that a threat actor produces. As the most visible aspect of an influence operation, content often draws the most public attention. However, analysing content without understanding the actors and behaviours behind it is insufficient.

The table below highlights the questions to ask in identifying content in an IMI incident:

Content variable	Question
Truthfulness	Is the content verifiably untrue or deceptive?
Narrative(s)	Does the content align with known disinformation narratives?
Language(s)	Which languages do the actors use in the spread of the disinformation or other online content in question?
Synthetic	Is the content manipulated or artificial?
Expression	Is the content reasonable self-expression protected by fundamental freedoms?
Harm	Is the content dangerous?

Real-life example:
Zimbabwe's 2018 general elections

EXAMPLE

To undermine his credibility, disinformation tactics targeted opposition candidate Nelson Chamisa in multiple languages, falsely claiming he would give away the country's natural resources if elected. Social media and encrypted apps such as WhatsApp amplified the falsehoods, making fact-checking difficult. Some content included doctored images and videos showing Chamisa with foreign officials. The campaign deepened political divisions and fuelled concerns over the line between political expression and voter manipulation.

Click here to read the ICFJ article.

Degree

This element examines how widely content spreads and the audiences it reaches in a given case. Assessing the scale of distribution helps decision-makers determine whether countermeasures are necessary. This component tracks hashtags, networks, shares, and other indicators of reach and engagement.

The questions to ask in identifying the degree of an IMI incident:

Degree assessment	Question
Audience(s)	Who constitutes the content's target audience(s)?
Platform(s)	Is it possible to map which channels or platform(s) actors use to distribute the content and interact with audiences?
Virality	Is the content going viral on social media platforms in a way that would suggest an inauthentic boost in engagement?
Targeting	Is the content tailored or micro-targeted, and, if so, to which audiences?
Scale	Does the scale of the incident indicate a single operation or an ongoing campaign?

Case study:
Nigeria's 2023 general elections

Disinformation campaigns targeted voters across social media platforms, particularly Facebook and X, using a mix of bots (automated accounts), cyborgs (hybrid accounts combining automation with human oversight), and trolls (individuals who deliberately provoke or disrupt discussions to amplify misleading narratives). These tactics artificially boosted engagement, drowning out legitimate discourse. The content was micro-targeted at specific voter demographics, particularly young and undecided voters, with tailored messaging designed to sway opinions. Exploiting the digital architecture of online platforms, disinformation spread rapidly, leveraging hashtags and coordinated inauthentic behaviour to maximise reach and virality. The sustained nature of these campaigns suggests a well-organised influence operation rather than isolated incidents.

Nigeria's 2023 elections saw a surge in coordinated disinformation, aimed at manipulating voters and discrediting candidates. Fabricated results, such as false claims of an APC victory in Lagos, and inauthentic tactics like mass copy-pasting and hashtag manipulation, distorted public perception. Social media influencers and bot accounts amplified these narratives, targeting marginalised voices and exploiting ethnic tensions. Disinformation was widespread across political parties, with false claims of Labour Party victories also circulating. The impact was significant, eroding trust in the electoral process, fuelling political tensions, and underscoring the urgent need for media literacy, proactive information management, and greater accountability to curb harmful misinformation.

Screenshot of former presidential aide Bashir Ahmad amplifying the fabricated results on X (Source CfA via HumAngle)

Effect

The effect component of the ABCDE framework assesses the threat level of a case by analysing its impact. Indicators, derived from the first four components, help determine the overall effects and inform response strategies.

The table below outlines the useful questions to ask in identifying the effect of an IMI incident:

Effect component	Question
Climate of debate	Is the online content issue-based? Does it, for example, involve false information, polarisation, or trolling?
Trust and reputation	Is the content target-based? Does it, for example, involve false rumours, cybersecurity hacks, forgeries and/or media leaks?
Fundamental freedoms	Is the content denying a fundamental freedom? For example, does it seek to deny freedom of expression or of political deliberation?
Public health	Does the content threaten individuals' health, medical safety, or physical wellbeing?
Public safety	Does the content threaten individuals' physical wellbeing or public order?
Election integrity	Does the content dissuade voters from participating in elections or seek to undermine the results of an election?
National security	Does the content threaten the territorial integrity or the national security of a sovereign state?

Real-life example:
Mali's 2020 legislative elections

EXAMPLE

The 2020 legislative elections in Mali took place in a politically volatile environment, marked by armed conflicts, military coups, and widespread anti-government protests. President Ibrahim Boubacar Keïta's administration, which was later overthrown in a coup that same year, faced mounting pressure from opposition groups and extremist organisations. Against this backdrop, disinformation campaigns became a powerful tool in shaping public perception and deepening political instability.

Various actors engaged in these campaigns, each seeking to manipulate the narrative to their advantage. Pro-government figures, including politicians and state-aligned media outlets, spread claims linking opposition parties to extremist groups such as Al-Qaeda and the Islamic State. At the same time, disinformation targeting opposition groups, such as the Rassemblement pour le Mali (RPM), portrayed them as responsible for escalating instability in northern Mali.

Local influencers, often with political motivations, amplified these narratives on social media, further polarising public sentiment. Digital platforms became battlegrounds for competing factions, where manipulated content, fake endorsements, and misleading claims worked to discredit political opponents and sway voter opinions. In a fragile democracy already weakened by security threats and governance crises, these disinformation efforts played a significant role in fuelling distrust and uncertainty around the electoral process.

3.2.2. Directing Responses Against Illicit Influence Operations (D-RAIL)

The D-RAIL framework counters illicit influence operations that undermine democratic institutions by disrupting their operational mechanisms. Instead of focusing solely on disinformation, it targets coordinated, covert campaigns led by geopolitical actors. By leveraging data-driven analysis, D-RAIL identifies weak points and applies strategic interventions to hinder illicit influence efforts. Its adaptable structure allows it to function across various platforms and formats, continuously evolving with new data to strengthen long-term defences.

To effectively implement the D-RAIL framework against illicit influence operations, follow these steps:

1. **Map the chain of influence:**
Identify elements such as actors, information environments, intended effects, and targets to understand how the operation functions.

2. **Disrupt the chain:**
Develop targeted interventions, such as exposing tactics, countering false narratives, or limiting reach through platform measures.

3. **Monitor and evaluate:**
Track both the influence campaign and the countermeasures, using data to assess their effectiveness.

4. **Adapt and refine:**
Continuously improve strategies based on insights from monitoring, ensuring countermeasures evolve with emerging threats.

5. **Maintain ethical standards:**
Adhere to ethical principles, build necessary skills, and ensure transparency to safeguard democratic processes.

4.

Information manipulation: What is it?

4. Information manipulation: What is it?

This chapter explains what information manipulation and interference (IMI) means and who is typically involved in pushing IMI campaigns. It covers what constitutes online harm from the perspective of various social media platforms. It also discusses the differences between misinformation, disinformation, and malinformation using real-life examples and case studies.

4.1. Understanding the IMI and FIMI landscape

Information manipulation and interference (IMI) refers to behaviour that threatens political processes or values. It can be locally driven or foreign-sponsored (FIMI), benefitting external actors.

Elements of IMI/FIMI

Actors:
Individuals, non-state entities, organisations, or state entities (e.g. extremist groups, influencers, media outlets, or politicians). Identifying actors helps track patterns and motives.

Narratives:
Structured messages influencing public perception.
These fall into:

i. **Meta-narratives:**
Broad themes, e.g. anti-Western sentiment.

ii. **Sub-narratives:**
Specific claims, e.g. a government being weak or corrupt).

Who are the actors involved in IMI and FIMI campaigns and why?

State actors

Governments and state-sponsored entities deploy IMI campaigns to shape public opinion, influence geopolitical dynamics, and control narratives. They use coordinated information operations, cyber tactics, and propaganda through diplomatic channels, state media, and digital platforms to achieve strategic objectives.

Political organisations

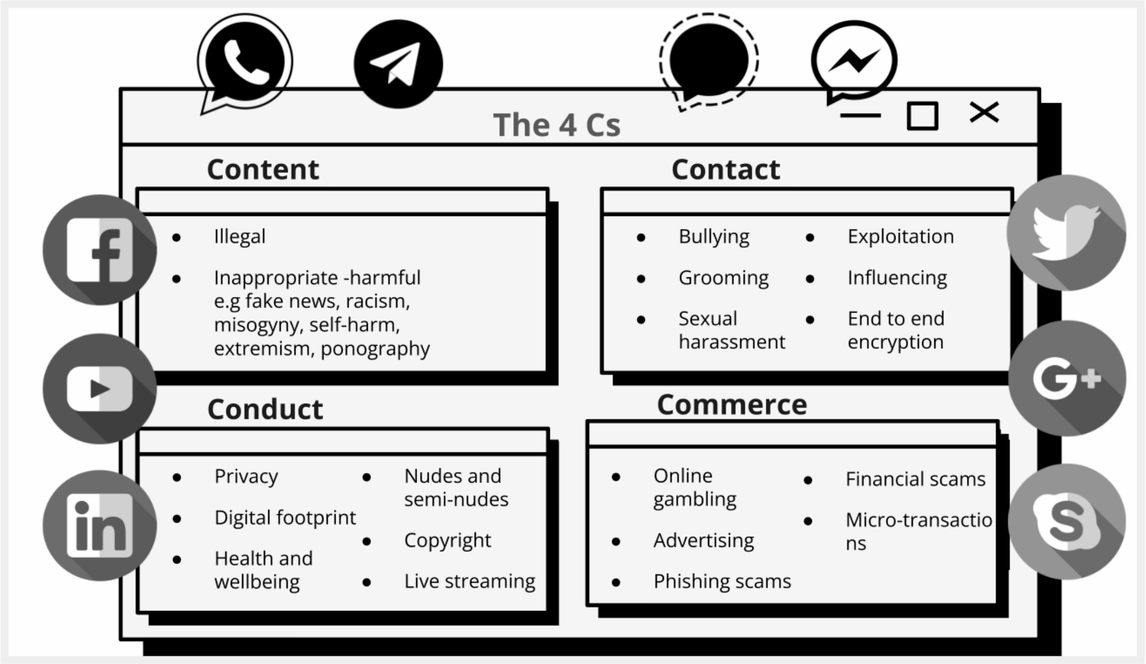
Advocacy groups, partisan organisations, and political parties use IMI tactics to discredit opponents, gain electoral advantages, and mobilise supporters. They exploit divisive issues, manipulate social media, and spread mis-/disinformation to sway public opinion.

Special interest groups

Companies, lobbyists, and NGOs leverage IMI tactics to influence policy, regulatory decisions, and public perception. They engage in astroturfing, advocacy campaigns, and media manipulation to promote their interests.

4.2. How digital platforms define ‘online harm’

To clarify the nature of social media harms and justify the need for policy action, experts developed the 4Cs framework, categorising various risks users face in digital spaces. These aim to define specific harms, their causes, and consequences, while broadening the conversation beyond dominant issues such as cyberbullying, grooming, and pornography.



4.2.1. Meta

Meta tackles online harm through its community standards and safety measures, though it does not explicitly define the term. Its approach focuses on content safety and user protection.
Meta divides online harm related to content into two categories:

i. **Content-related harm:** This includes violent and graphic content, dangerous organisations and individuals, hate speech and discriminatory content, child exploitation material, promotion of suicide and self-injury, and misinformation that could cause real-world harm.

ii. **Interaction-related harm:** This includes coordinated inauthentic behaviour, account compromise, platform manipulation, identity-based harassment, privacy violations, and voter interference and civic harm.

Meta's approach to online harm prevention is based on its safety documentation, which includes the following principles:

a. Balancing free expression with safety considerations.
b. Preventing harmful content and behaviour across its platforms.
c. Applying special protections for vulnerable groups, including minors, at-risk individuals, public figures, journalists, and marginalised communities.
d. Safeguarding civic integrity and preventing election interference.

Enforcement mechanisms

Meta uses a variety of mechanisms including AI detection and review systems working alongside human content moderation teams to identify and address harmful content. The platform implements content warnings and reduced distribution for borderline material while applying account-level interventions when necessary. Meta has also established a formal appeals process for users who believe their content was incorrectly removed, supplemented by an independent Oversight Board that reviews complex cases and makes binding decisions.

4.2.2. TikTok

TikTok does not explicitly define online harm. However, it addresses various forms of online harm through its community guidelines and safety measures. Here are the different dimensions of online harm that TikTok addresses:

a. **Content safety:** Online harm in the context of content includes two types of harm.

i. **Content-related harm:** Includes content that exploits or endangers minors, explicit content that violates community standards, graphic violence or dangerous challenges, and hate speech and discriminatory content.

ii. **Interaction-related harm:** Includes coordinated harassment campaigns, cyberbullying, predatory behaviour, sexual harassment, and targeted harassment.

b. **User protection principles:** Based on TikTok's trust and safety documentation, online harm prevention includes blocking content that could cause physical or psychological damage, preventing the spread of mis-/disinformation, protecting vulnerable user groups, especially minors, and safeguarding user privacy and personal information.

Enforcement mechanisms

These include AI-powered content moderation, human review teams, penalties for policy violations, proactive detection of potential harmful content, and reporting mechanisms for users.

4.2.3. Telegram

Telegram's approach to online harm differs from other platforms, emphasising user privacy and minimal central moderation.

The dimensions of online harm on Telegram are:

a. **Platform philosophy:** Prioritises free communication with limited content moderation, relying on user-level controls to mitigate harm.

b. **Content-related harm:** Includes direct calls for violence, explicit sexual content involving minors, extreme graphic violence, systematic harassment campaigns, and terrorist propaganda.

c. **Enforcement approach:** Enforcement is divided into two main categories:

   i. **User-level protection features:** Blocking/reporting mechanisms, message restrictions, personal information concealment, and privacy controls.

   ii. **Platform-level protection features:** Include decentralised moderation, emphasis on individual user choice and control, limited proactive content filtering, and user-driven reporting systems.

4.2.4. X (formerly Twitter)

X addresses online harm through algorithmic detection, human moderation, and a graduated penalty system, from temporary restrictions to permanent bans. Enforcement follows ethical, legal, and regulatory standards.

X has several categories for online harm:

a. **Abuse/harassment:** It prohibits sharing abusive content and engaging in targeted harassment.

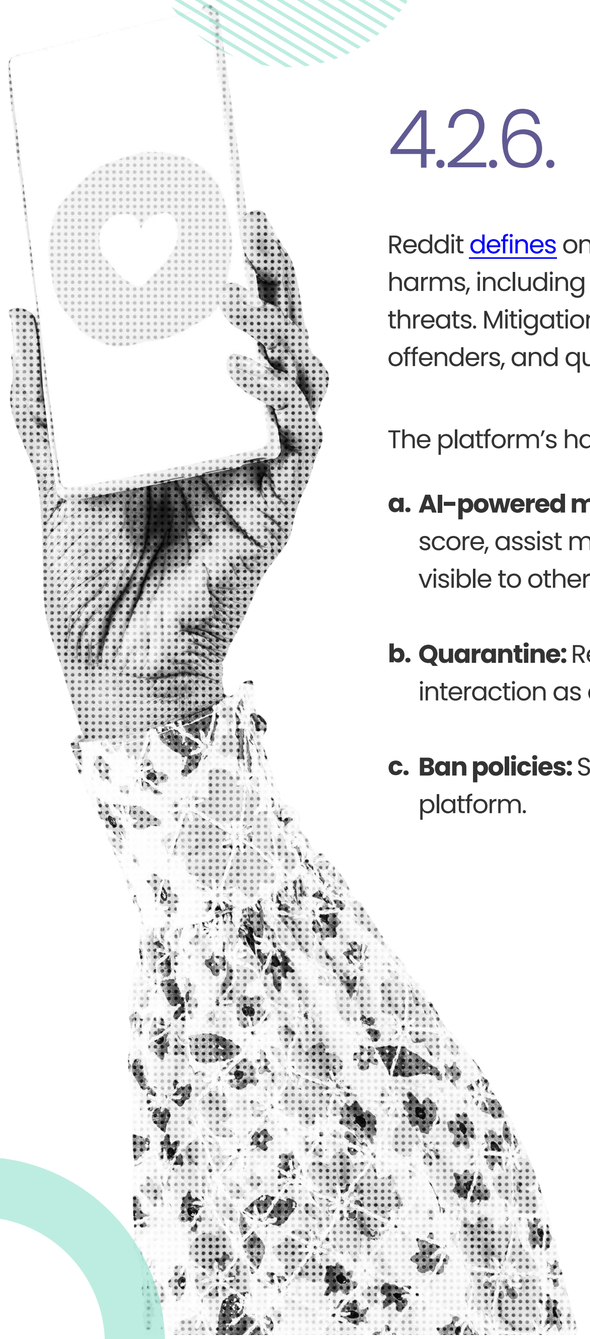
b. **Hateful conduct:** It defines hateful conduct as behaviour that is discriminatory, harassing, or intimidating.

c. **Violent speech:** The platform has softened its policy on violent speech, moving from a zero-tolerance approach to one where violent speech may be removed or have its visibility reduced. X says this is to balance free expression with user safety.

d. **Misinformation and extremist views:** X has features to block and report such content, but these are not always effective across all languages.

4.2.5. BlueSky

Bluesky addresses online harm through decentralised, community-driven moderation. Users can customise content governance, filtering harms including misinformation, harassment, and hate speech. Built on the authenticated transfer (AT) protocol, the platform offers open-source tools and independent moderation services. Ozone, BlueSky's collaborative moderation tool, enables content review and labelling, allowing users to refine their experience.



4.2.6. Reddit

Reddit defines online harm through prohibited content and interaction-related harms, including coordinated abuse, doxxing, hate speech, harassment, and violent threats. Mitigation strategies include AI-powered moderation tools, banning severe offenders, and quarantining rule-breaking communities.

The platform's harm mitigation strategies include:

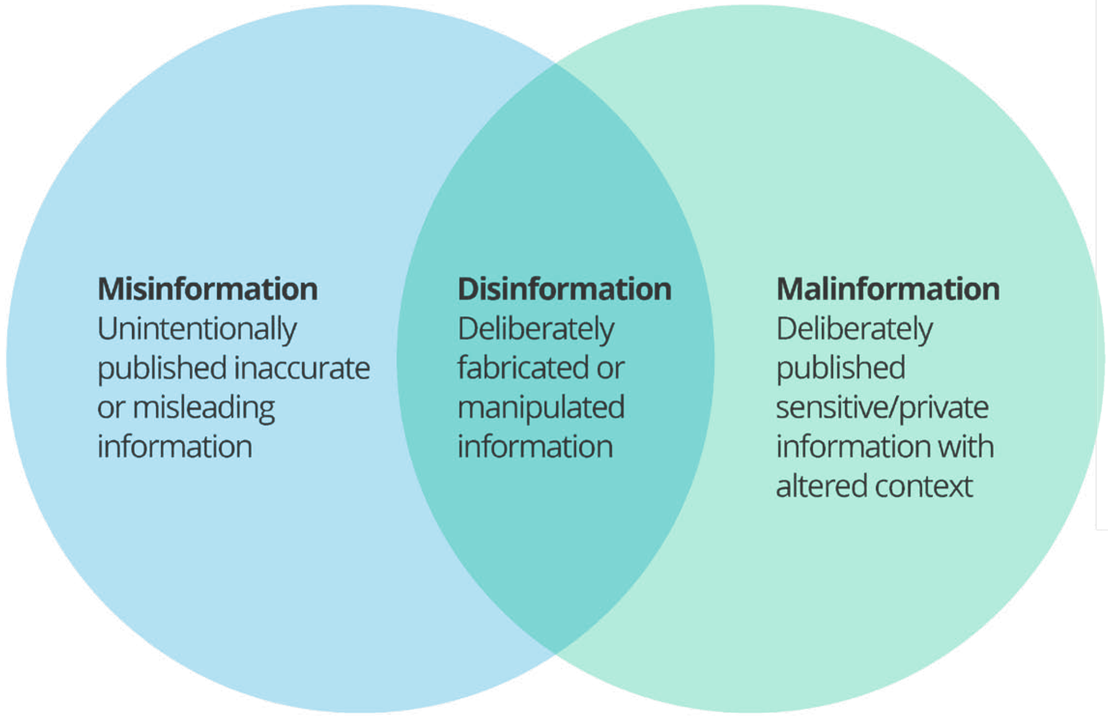
a. **AI-powered moderation tools:** AI-driven features, such as the contributor quality score, assist moderators in identifying high-quality contributors, though they are not visible to other Reddit users.

b. **Quarantine:** Repeated rule-violating communities face restricted visibility and interaction as a preliminary measure before a ban.

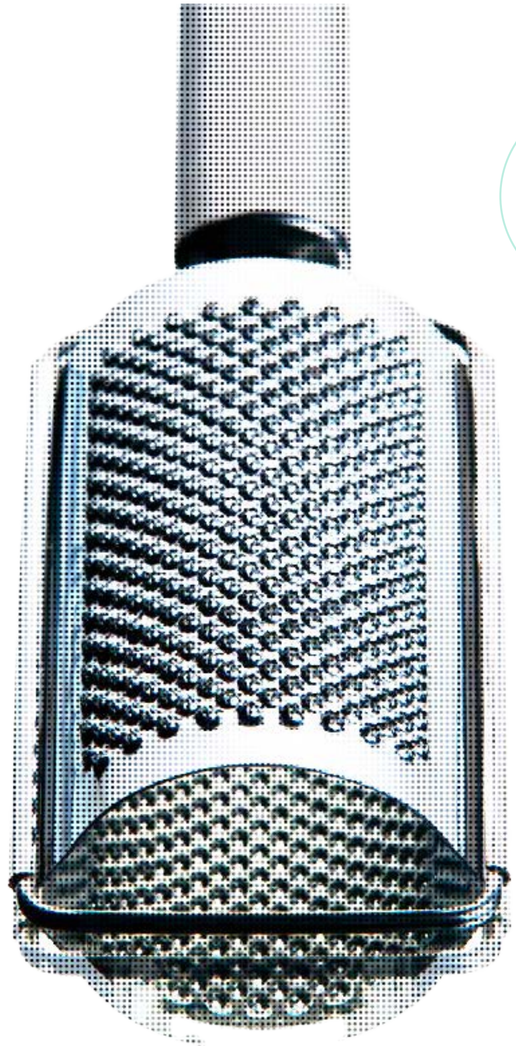
c. **Ban policies:** Severe or repeated violations lead to permanent removal from the platform.

4.3. **Mis-/dis-/mal-information**

In this diagram, it shows that some misinformation might unintentionally spread harmful effects similar to malinformation, and certain disinformation might be based on facts but twisted to mislead, blending elements of both dis- and misinformation.



A screengrab of an example showcasing the mapping of DRC regions and provinces
(Source: CfA using the DRC lexicon database)



4.4. Coordinated inauthentic behaviour

Coordinated Inauthentic Behaviour (CIB) refers to deliberate efforts by individuals or groups to manipulate public opinion through fake or misleading accounts, groups, or pages on social media. These networks operate covertly to deceive audiences and amplify specific messages for financial, political, or social influence.

For example, during an election, a coordinated network might create hundreds of fake accounts to spread false information about opposing candidates while posing as independent grassroots supporters. These accounts share fabricated news, amplify divisive rhetoric, and spread disinformation about voting procedures. By engaging with real users, they create the illusion of widespread support, artificially shaping public perception and influencing electoral outcomes.

Characteristics of CIB

a. **Coordination:** Multiple accounts or entities work together to promote certain actors, agendas, or narratives in a concealed and systematic way.

b. **Inauthenticity:** Automated bots, fake profiles, or pages that falsely present as independent users or organisations are used to spread mis- or disinformation or propaganda.

c. **Intent to deceive:** The purpose is to mislead the audience into believing the activity is organic engagement and shows genuine grassroots support for or against a cause.

d. **Manipulation:** It intends to disrupt discussions, influence public opinion, or shape perceptions in a way that benefits the actors or their agendas.

e. **Scale and impact:** Campaigns are designed to reach a wide audience.

f. **Evolving tactics:** Actors adapt their tactics to evade detection and counter platform defences, making it a persistent and evolving threat.

Different types of CIB

Posts: Multiple accounts broadcast the same message, seemingly independent of one another, yet they are managed by the same person or team.

Reposting and sharing: Fake or astroturfing accounts amplify a campaign message by reposting or sharing content.

Hashtags: Multiple accounts use the same hashtag within a specific time threshold to push a narrative.

Mentions and tags: Multiple accounts mention or tag the same user or entity to inject a disinformation narrative.

Comments: Multiple accounts generate comments to manipulate the discussion around a post or topic.

Engagement: Accounts perform synchronised actions such as liking, viewing, or reacting to content to boost its visibility.

4.4.1 Platform-specific definitions and actions against CIB:

In addressing cases of CIB, it is crucial to consider which platforms were used for the activity, and what the platform's policy is regarding it. This means that you can appeal to the definitions and regulations of the affected platform when mapping and reporting your findings.

Meta (Facebook, Instagram, and Threads):

On Meta, CIB is classified as groups of pages or individuals that collaborate to conceal their identity or activities, often using fake accounts. Meta targets networks that manipulate public discourse for strategic purposes with actions including:

a. Removing fake accounts, pages, and groups involved in coordinated manipulation
b. Conducting investigations based on tips and automated detection
c. Publishing regular reports on CIB takedowns
d. Working with third-party fact-checkers to identify and label misinformation

X (formerly Twitter)

X focuses on '[platform manipulation and spam]', which includes CIB designed to manipulate public perception or disrupt user experience, often through bots and fake accounts influencing trends and discussions.

Actions

i. Suspending accounts engaged in platform manipulation and spam.
ii. Removing or labelling content that violates policies.
iii. Implementing measures to limit the spread of mis-/disinformation.

Telegram

CIB on Telegram shows as coordinating bots, channels, and groups to spread propaganda and manipulate public opinion. This includes the rapid sharing of media, messages, and links across multiple groups.

Actions

i. Relying on user reports.
ii. Taking action against channels or groups that violate terms of service, especially those linked to illegal activities or spam.

TikTok

On TikTok, CIB [manifests] as the coordinated creation and promotion of misleading content, often leveraging challenges, sounds and trends to manipulate views and influence public sentiment. This includes using fake accounts and bots to artificially boost video views and engagement.

Actions

i. Using automated detection and human review to identify and remove inauthentic content and accounts.
ii. Taking action against accounts involved in coordinated manipulation, such as terminating the accounts or removing their content.
iii. Collaborating with fact-checkers to identify and label mis-/disinformation.
iv. Running public-facing education campaigns.
v. Publishing transparency reports.

Reddit

CIB on Reddit often involves the coordinated manipulation of subreddits and the upvoting/downvoting system to promote specific agendas or narratives. This can include creating fake accounts, using bots, and coordinating content posts to influence discussions and trends.

Actions

i. Banning accounts and subreddits involved in coordinated manipulation, particularly those violating content policies or manipulating the platform's voting system.
ii. Relying on user reports and automated detection to identify and remove inauthentic activity.
iii. Using moderators for each subreddit to help maintain the integrity of content.
iv. Taking action against brigading, a form of CIB.

WhatsApp

On WhatsApp, CIB is seen as the coordinated spread of false information and propaganda through group chats and forwarded messages. This can include the use of automated bots and the manipulation of media to deceive users.

Actions

i. Limiting the forwarding of messages to reduce the spread of viral misinformation.
ii. Suspending accounts engaged in spamming or automated messaging.
iii. Encourage users to report misinformation.

4.5. Meta’s approach to tackling hate speech

Social media platforms define and address hate speech differently. This section explores Meta’s approach, investigative strategies, and a case study from Kenya’s 2022 elections.

Meta's transparency guideline on hate speech outlines parameters without providing a strict definition, indicating that hate speech includes attacks or demeaning content based on protected characteristics like disability, ethnicity, gender, race, religion, or sexual orientation, as well as content using images, text, or videos to target specific groups, and the use of idioms, hidden meanings, nuances, code words, misspellings, or symbols to convey hateful messages while evading detection.

EXAMPLE

Real-life example:
Hate speech in Kenya during 2022 elections

During the August 2022 Kenyan elections, senator Mithika Linturi used the Swahili word ‘madoadoa’, [‘spots’ or ‘blemishes’ in English]. However, this term is also an ethnically coded slur. Linturi used madoadoa to call for the removal of ‘outsiders’, referring to them derogatorily as ‘spots’ within the community. His remarks targeted ethnic communities that had permanently settled in the Rift Valley but were not indigenous to the region.

The National Cohesion and Integration Commission (NCIC) classified ‘madoadoa’ as hate speech due to its inflammatory nature and potential to incite violence.

Screenshot highlighting the use of the word ‘madoaodoa’
(Source: CfA via Tuko)

4.5.1. Meta's 'market-specific slur lists'

According to Meta's market slur community standards, 'a slur is a word that is inherently offensive and used as an insult toward a protected characteristic.

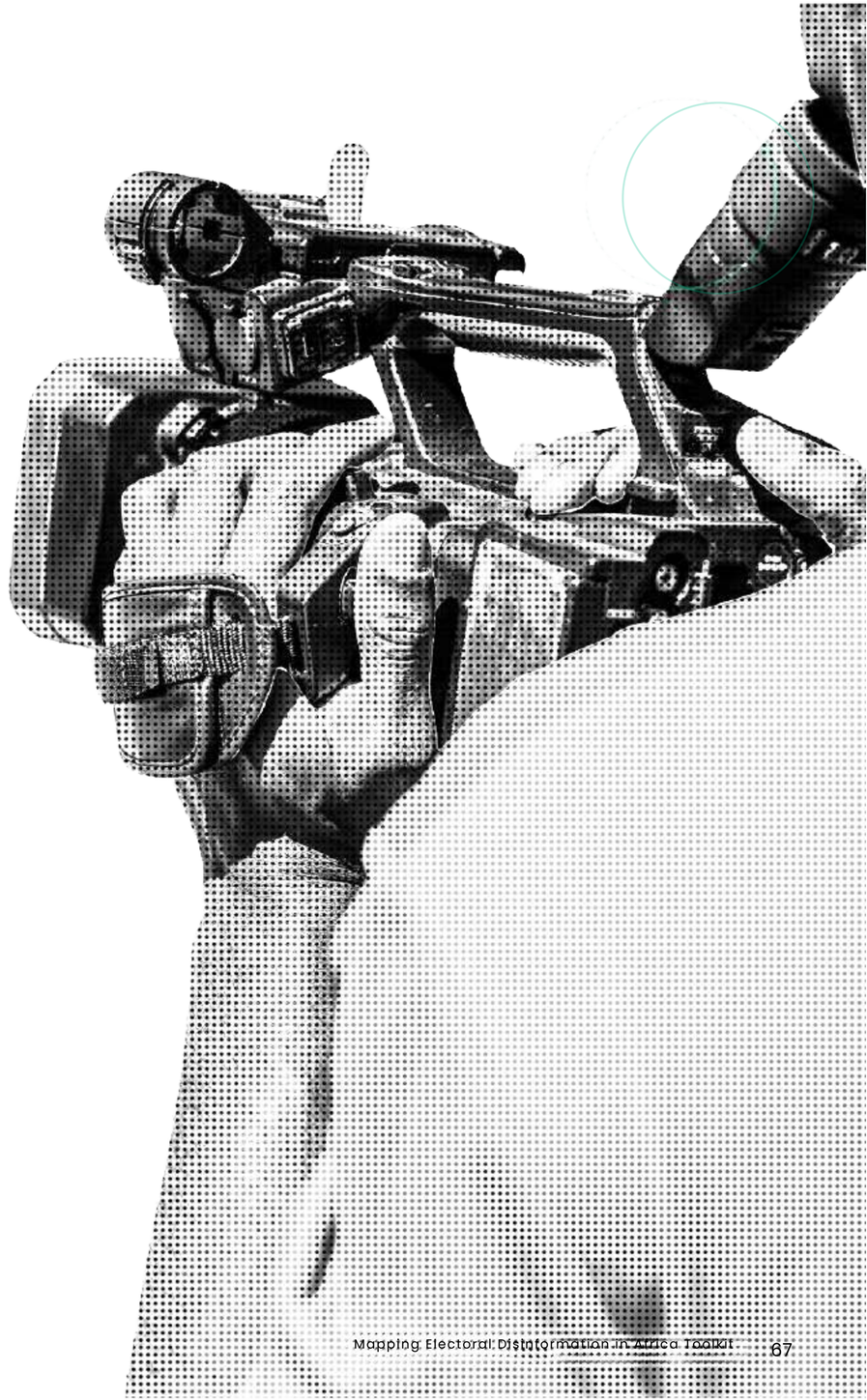
How to apply this information to investigate on Meta platforms

An understanding of market slurs can be applied to investigate individuals and groups on Meta platforms such as Facebook and Instagram:

a. **Know Meta's policies:** Slur lists are market-specific, as slurs vary by region.
b. **Report slurs:** Flag content that demeans or insults individuals or groups.
c. **Track variations:** Identify evolving slurs and report them to improve Meta's detection.

4.6. Suppression/censorship + mass reporting

As social media platforms expand to accommodate millions – or even billions – of users, they increasingly delegate content moderation to automated systems. These systems help detect issues like spamming and impersonation but are vulnerable to manipulation and misuse.



4.6.1. Meta mass reporting

Mass reporting is the coordinated misuse of platform reporting systems to silence content, individuals, or groups. It relies on volume-based abuse of moderation tools to suppress voices, often without legitimate cause.

How this work

a. **Coordination:** Organised efforts, via social media or private messaging, encourage mass reporting.
b. **Exploiting moderation systems:** Platforms with automated or semi-automated moderation may take action based on report volume rather than content merit.
c. **Impact:** High report volumes can result in account suspensions, content removal, or reduced post visibility.

Intent and weaponisation

a. **Silencing dissent:** Used to suppress inconvenient truths, opposing views, or unpopular opinions.
b. **Digital censorship:** Circumvents traditional oversight, allowing groups to enforce their own speech standards.
c. **Targeting marginalised voices:** Activists, journalists, and vulnerable communities often face mass reporting as a tool of suppression.

4.7. Coordinated brigading/doxxing/trolling

Coordinated online manipulation uses tactics such as brigading, doxxing, and trolling to influence public opinion and suppress dissent. These methods spread disinformation and create intimidation. The Bell Pottinger/Gupta scandal demonstrates the real-world impact of such manipulation.

Real-life example:
Bell Pottinger/Gupta scandal

Screenshot of the story on the Times Live SA((Source: CfA via Youtube)

The Gupta family, closely linked to former South African president Jacob Zuma, faced allegations of 'state capture'. To counter these claims, the family hired PR firm Bell Pottinger, which crafted the 'white monopoly capital' (WMC) narrative, exploiting racial tensions to deflect corruption accusations. Bell Pottinger orchestrated online campaigns using fake accounts to attack critics, and dox opponents, and spread the WMC message, creating an echo chamber that fuelled division.

The WMC narrative deepened racial tensions, while online harassment silenced dissent. The backlash ultimately led to Bell Pottinger's collapse, contributed to political instability, and played a role in Zuma's removal as president. This case underscores the dangers of online manipulation and the ethical responsibility of PR firms and social media platforms.

Click here to read ANCIR's investigation into WMC.

4.8. Blackmail/hacking/surveillance

The rise of digital surveillance technologies has become a growing concern in many countries, particularly during election periods, where governments have been accused of exploiting these tools to silence dissidents and gain unfair political advantages. Reports by the non-profit Privacy International indicate that unauthorised access to personal data, combined with advanced surveillance capabilities, has enabled the blackmail, profiling, and tracking of activists, critics, and political opponents.

These practices undermine democratic processes, erode public trust, and highlight the need for accountability, transparency, and legal safeguards to prevent the misuse of surveillance for political manipulation.

4.8.1 Meta cyber espionage

Cyber espionage involves covert digital tactics to access confidential information from governments, individuals, or organisations, often through state-sponsored data exfiltration, hacking, or surveillance, for economic gain, intelligence or political control.

Real-life example:
Surveillance and unauthorised cyber operations in Kenya

EXAMPLE

In Kenya, reports indicate that intelligence and counterterrorism agencies, in collaboration with telecom companies, have engaged in unauthorised cyber operations, including accessing customer data without due process and using foreign surveillance technologies. These practices threaten electoral integrity by enabling profiling, blackmail, and tracking of political opponents, eroding public trust and violating privacy rights. Safaricom has denied involvement, while the government defends its actions as lawful, but concerns persist due to weak oversight and potential misuse of surveillance for political ends.

Click here to read the Quartz article.

4.9. Illicit Influence

Illicit influence refers to any intentional actions or activities designed to disrupt, manipulate, or undermine the integrity of an electoral process. These include, but are not limited to, fraud, manipulation of voting procedures, micro-targeting voters, threats, or the use of social media and other technologies to spread disinformation or unduly influence voter behaviour.

The following are some common tactics used for illicit influence in elections:

Influence for hire

This is the practice of outsourcing public opinion manipulation to third-party entities. Using data analytics, micro-targeting, and social media campaigns, these services alter perceptions, shape narratives, and sway political outcomes. They also strategically deploy tailored messaging or disinformation.

Influence for hire can serve multiple purposes, including:

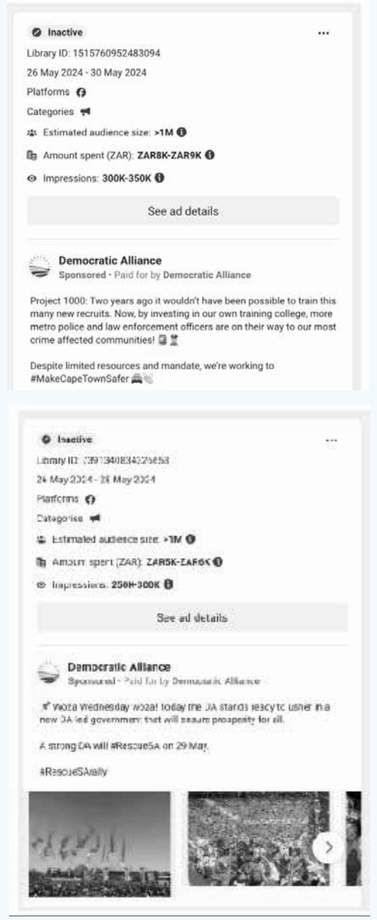
i. **Election manipulation:** Swaying voter opinions through targeted content and ads.

EXAMPLE

Real-life example:
Opaque political ads in South Africa

In the lead-up to South Africa's May 2024 elections, political campaigns in the country relied on micro-targeted digital ads on platforms such as Facebook and Instagram. The Democratic Alliance spent more than R2 million (US$110,737) on 799 ads, while Freedom Front Plus and Rise Mzansi spent R250,000 (US$13,839) and R50,000 (US$2,767), respectively. Nonprofits such as Ask South Africa and Pledge to Vote SA also ran targeted ads at R233,000 (US$12,893) and R46,000 (US$2,544), respectively, often advancing political agendas such as persuading users to vote for certain political parties. Notably, organisations such as Ask South Africa and Pledge to Vote SA did not disclose their affiliations with political parties or the sources of their funding, making it unclear who was behind the messaging and whether it was politically motivated.

The lack of transparency in these efforts raises concerns about election manipulation, as tailored and undisclosed messaging reinforces echo chambers and influences voter opinions.



Sample political ads running in May 2024
(Source: CfA via Meta Ads Library)

ii. **Reputation management:** Destroying or protecting the reputation of governments, individuals, or organisations through online campaigns.

EXAMPLE

Real-life example:
Paid influencers in Nigeria

A BBC investigation found that Nigerian political parties paid social media influencers to spread false stories ahead of the 2023 elections, including claims linking opponents to extremist groups. In exchange for cash, gifts, or political positions, influencers amplified these narratives, which often spread offline, further eroding trust in the electoral process. Despite being illegal, such tactics persist with little oversight.

Click here to read the BBC article.

iii. **Propaganda:** The spread of information, whether facts, arguments, rumors, half-truths, or falsehoods, intended to sway public opinion.

EXAMPLE

Real-life example:
Propaganda affecting 2024 US election

In the 2024 U.S. presidential election, Chinese and Russian state-run media used selective coverage as a form of propaganda. Chinese outlets such as Xinhua and the Global Times, provided limited coverage of the Harris-Trump debate, avoiding positive portrayals of US democracy, while focusing on Biden's previous debate to highlight democratic flaws. Meanwhile, Russian media, for example RT and Sputnik, downplayed Harris's performance and subtly favoured Trump, reflecting Russia's alignment with his policies. Both nations strategically used media to shape public perception and influence the election narrative.

Click here to read the VOA article.

Bot networks and fake accounts

Both bot networks and fake accounts can significantly impact elections by:

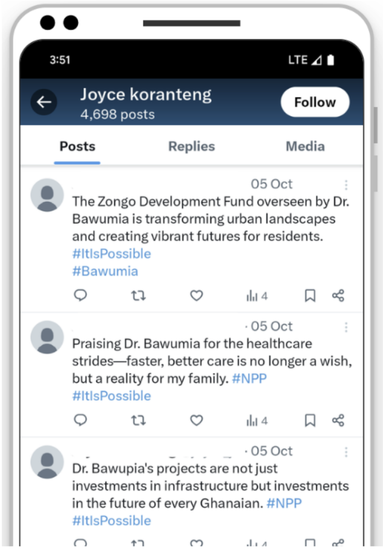
i. **Manipulating public opinion:** By flooding social media with specific messages or hashtags, they create the illusion of widespread support for a candidate or ideology, swaying undecided voters.

EXAMPLE

Real-life example:
Bot accounts influencing Ghana election

Ahead of Ghana's 07 December 2024 presidential election, 171 bot accounts on X used AI-generated content to promote incumbent president Mahamudu Bawumia and discredit opposition leader John Mahama. These bots amplified pro-Bawumia narratives and right-wing messaging, pushing hashtags such as #Bawumia2024 to manufacture the illusion of mass support. Their predictable posting patterns and reliance on AI tools such as ChatGPT suggest a coordinated effort to manipulate public opinion and sway undecided voters, exacerbated by weaker content moderation on the platform since Elon Musk's acquisition.

Click here to read the Rest of World article.



An example of a phony X account meddling in Ghana's election (Source: CfA via Rest of World)

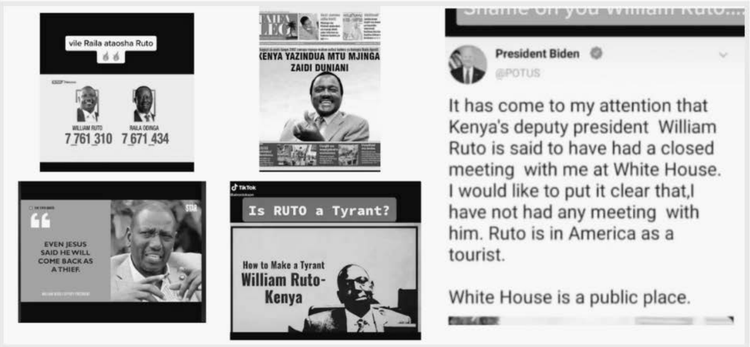
ii. **Spreading disinformation:** They can spread biased, false, or misleading information about political candidates, issues, or policies, which can distort the electorate's perception and undermine trust in the democratic process.

iii. **Undermining the credibility of opposition:** Fake accounts often engage in smear campaigns against political opponents, attempting to discredit their credibility and reputation.

Real-life example:
Political disinformation in Kenya

EXAMPLE

Ahead of Kenya's 2022 general election, there was a rise in political disinformation on TikTok, a platform that gained influence despite its policies against political content. A study by the Mozilla Foundation examined over 130 videos from 33 accounts, revealing widespread use of hate speech, incitement to violence, and manipulated content, with TikTok's algorithm often amplifying these videos. These videos, which included ethnic threats and false narratives about candidates, received over four million views, contributing to the polarisation of voters. The platform's failure to effectively moderate such content, compounded by issues including context bias and rushed moderation, made TikTok a significant source of disinformation during the election period, raising concerns about its impact on young, impressionable voters.



Screenshots of examples of synthetic content circulating on Tiktok
(Source: CfA via the Mozilla Foundation)

Click here to read the Mozilla Foundation article.

Case study:
Fake accounts influencing US election

A report by Cyabra, an Israeli tech company, revealed that fake accounts on X had increased ahead of the 2020 US presidential election, with 15% of pro-Trump accounts and 7% of pro-Biden accounts identified as inauthentic. The pro-Trump accounts were part of a coordinated campaign, using tactics including synchronised posts and identical hashtags such as 'Vote for Trump' and 'Biden is the worst president the U.S. has ever had', to promote negative narratives about Biden. This manipulation undermined credibility by amplifying harmful messages, influencing public perception, and skewing online discussions in Trump's favor. In contrast, fake pro-Biden accounts showed no signs of coordination.

Click here to read the Reuters article.

Deepfakes and manipulated media

Deepfakes and manipulated media threaten election integrity by spreading false information and misleading voters. AI-generated deepfakes create hyper-realistic but fake videos or audio clips that falsely attribute statements or actions to political candidates, damaging reputations and swaying public opinion. Similarly, doctored images and videos are often deployed to mislead or incite unrest.

During elections, these tactics erode trust in candidates and electoral processes, fuel disinformation campaigns, and deepen political polarisation.

Real-life example:
Deepfake of Chinese president

EXAMPLE

In December 2023, a deepfake video of Chinese president Xi Jinping began circulating on TikTok, appearing to encourage Taiwanese citizens to vote. Originally created as a parody, the video was later repurposed with altered captions supporting both the Kuomintang (KMT) party and Taiwan People's Party (TPP), misleading audiences about Xi's stance.

Similarly, in March 2024, a deepfake of US president Donald Trump was used to endorse uMkhonto weSizwe (MK), a new political party in South Africa, ahead of the country's May 2024 elections. The video gained traction on WhatsApp and X, spreading among MK supporters before fact-checkers debunked it.

While TikTok and X removed some flagged content, enforcement remained inconsistent, allowing manipulated videos to reach thousands before takedown.

Click here to read the Record Future article.

Screenshots of the deepfakes of Xi (left) and Trump (right)
(Source: AFP Fact Check and Taiwan FactCheck Center)

Influence through state-sponsored media

This refers to how governments leverage state-controlled or aligned media to shape public opinion, manipulate narratives, and advance political agendas, particularly during elections. By discrediting opposition, promoting government policies, and silencing dissent, these media outlets create an uneven playing field. The lack of diverse perspectives restricts informed decision-making, ultimately undermining democratic integrity and fair competition.

Case study:
State media bias in Zimbabwe

In the 2018 Zimbabwe presidential elections, state-controlled media, particularly the Zimbabwe Broadcasting Corporation (ZBC), played a crucial role in promoting the ruling Zimbabwe African National Union – Patriotic Front (ZANU-PF) party and president Emmerson Mnangagwa while marginalising opposition voices. The broadcaster's pro-government bias was evident as it provided disproportionate airtime to the ruling party and heavily criticised opposition politicians, including Chamisa from the Movement for Democratic Change (MDC) Alliance. This media manipulation, combined with contested postal voting, irregularities in voter rolls, and reports of voter intimidation, raised concerns about the fairness of the election.

Click here to read the Freedom House article.

4.10. FIMI vs IMI

IMI activities may not necessarily be illegal but are manipulative in character, and are conducted in an intentional and coordinated manner.

Real-life example:
Disinfo campaign against US presidential candidate

EXAMPLE

On 17 September 2024, a Microsoft research revealed operations by Kremlin-aligned troll farm, 'Storm-1516' on the same month to spread disinformation against America's presidential candidate Kamala Harris alleging she had left a 13-year-old girl paralyzed after an alleged hit-and-run in San Francisco in 2011. The video containing the claim was seeded by a fake website for a non-existent San Francisco news outlet named 'KBSF-TV'. The video was subsequently spread by X.com assets linked to 'Storm-1516', using the hashtag #HitAndRunKamala.

According to the research,an actor was paid to appear as the alleged victim in the video. According to the research, the website KBSF-TV was created shortly before the publication of the first related article about the alleged driving incident, per its online registration records. The video was aimed at discrediting Harris and causing controversy during her campaign. Storm-1516, which is part of Russia's propaganda operation, is known for using fake videos, audios, photographs and documents to back false claims on various social media platforms.

One such account, though currently suspended, that shared the video on X on 03 September 2024 is Aussie Cossack. According to the research, the account described itself as a 'Registered foreign agent for Sputnik News.' It shared the video with the message 'make this go viral MAGA folks'. Notably, CfA had previously flagged an account with a similar name on Telegram as having funded pro-Russian protests in Ghana in August 2023, days after the Nigerien coup.

Click here to read the Microsoft threat report article.

4.11. Media capture

External influence peddlers are increasingly compromising African media independence, leading to media capture. This occurs when governments, political elites, or lobby groups manipulate editorial policies and financial dependencies to control narratives. This report focuses on content manipulation – where external actors infiltrate newsrooms through paid punditry, content-sharing agreements, or sponsored journalism training. A common tactic involves foreign-backed pundits writing op-eds without disclosing financial ties, which state-controlled media then recycle as 'news'. This practice turns opinion analysis into a tool for propaganda, serving foreign or corporate interests over public discourse.

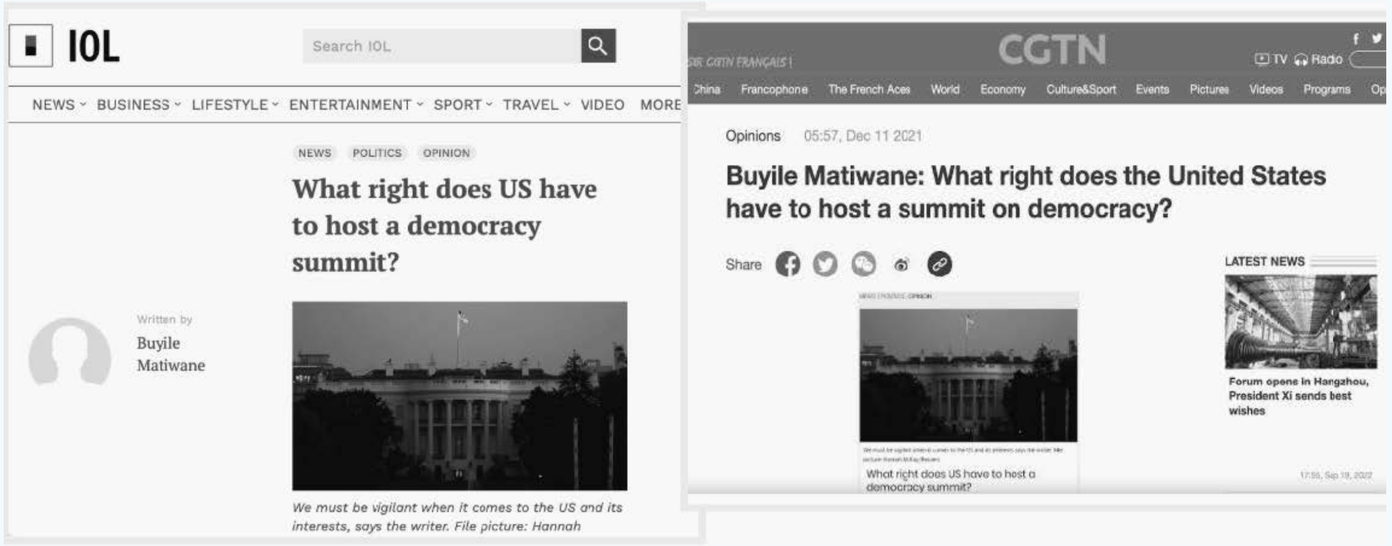
EXAMPLE

Real-life example:
Rinse cycling of a propagandistic pundit in South Africa

A South African student activist and budding politician, Buyile Matiwane (now deceased), was a politician in South Africa. He served as the South African Student Congress (SASCO) deputy president and was an African National Congress Youth League member. He wrote op-eds published in local media in support of the People's Republic of China (PRC), which then appeared in Chinese state media as news.

In total, Matiwane wrote 17 articles in defence and support of the PRC and published them on Independent Online (IOL), a South African news media site. These articles covered topics such as China-South Africa relations, PRC's foreign policy in Tibet, BRICS cooperation, and criticism of US democracy, Western media bias and PRC's president Xi Jingping's leadership.

On 08 December 2021, IOL published Matiwane's op-ed which said that a planned US summit on democracy was hypocritical. The French-language website of the Chinese state-run news channel CGTN then published a news story on 11 December 2021 reporting that South African outlets Pretoria News, The Cape Times, and The Mercury had published an opinion article authored by Matiwane. It is the organisational practice of the Independent Media Group, which operates all the listed outlets, to republish content across its network. The CGTN article summarises Matiwane's opinion piece, quotes paragraphs, includes a screenshot of the IOL article, and links to the original piece. Although the CGTN article refers to Matiwane as the author, it does not disclose his affiliations.



Screenshot showing Matiwane's op-ed as published by IOL on 08 December 2021 and an edited version republished by CGTN on 11 December 2021 (Source: CfA via IOL and CGTN)

On 11 March 2022, Chinese state-run news agency Xinhua reported on another of Matiwane's op-eds in French. IOL originally published the article, which discussed the US record on human rights, on 09 March 2022. Matiwane also weighed in on the Beijing Winter Olympics, writing on IOL on 27 December 2021 that the games should not be a platform for political posturing. Xinhua incorporated quotes from this op-ed into a news story, which it published on 01 January 2022.

These op-eds follow Matiwane's tenure as a 2019 Dongfang scholar at Peking University in Beijing in 2019, where he studied governance and policy. He spent six months in China as part of the exchange programme. In October 2020, Matiwane delivered a speech to provincial leaders in PRC, which was published on the SASCO Facebook page. According to the Facebook post, Matiwane had been in China for a number of weeks, alongside 'presidential advisors, journalists, academics, policy heads, kings, chiefs of staff, data analysts, career diplomats, and military men' for a programme about cooperation with China.

5.

Mapping and monitoring news media: Building an early warning system

5. Mapping and monitoring news media: Building an early warning system

Understanding a country's media ecosystem is essential for tracking information flow and detecting manipulated content. News monitoring acts as an early warning system, helping identify shifts in public sentiment and threats to electoral integrity. This section introduces CivicSignal's MediaData and MediaCloud, tools for mapping and monitoring digital news media. It covers narrative tracking, lexicon building, and integrating human intelligence to strengthen early detection systems.

5.1 Media mapping

MediaData maps media across the continent by profiling media and media-related organisations, media owners and industry professionals using a schema built from schema.org.

It also tracks media regulators, ombuds, support organisations, and training institutions. Researchers update the database through surveys and desktop research, sourcing data from regulatory bodies, and in-country experts. The mapping also includes organisation and industry professionals' social media links and size to understand their influence and reach.

	Organisation Name	Website	Audien...	Editorial ...	Editorial F...	Primary M...	Social: Facebook URL	Social: LinkedI...
1	1 FM / One FM	https://www.ac...	Municipal	Newsroom	General News	Radio	https://www.facebook.co...	
2	009 TV	https://www.ac...	Municipal	Newsroom	General News	Television	https://www.facebook.co...	
3	a ChaCooky Production	https://www.ac...	National	Production	Entertainment	Online	https://www.facebook.co...	
4	A24 (Africa 24) Media	https://a24me...	Global	Production	General News	Online	https://www.facebook.co...	https://www.linkedin...
5	Abagusii Global Radio	abagusiiglobal...	National	Newsroom	General News	Radio	https://web.facebook.co...	https://www.linkedin...
6	Ability Africa Magazine	http://abilityafri...	Sub-natio...	Newsroom	General News	Print	https://www.facebook.co...	https://www.linkedin...
7	Ability TV	https://www.thi...	National	Newsroom	Gender	Television	https://www.facebook.co...	https://www.linkedin...
8	ABP (Agence Burundaise de...	https://abpinfo...	National	Newsroom	General News	Online	https://www.facebook.co...	
9	Ace TV	http://acetv.co....	National	Newsroom	General News	Television	https://www.facebook.co...	
10	Acious Media	https://acious...	National	Production	Science & Tec...	Online	https://web.facebook.co...	https://www.linkedin...
11	ADMS (Africa Digital Media ...	https://adms.c...	Regional	Production	General News	Television	https://facebook.com/ad...	https://www.linkedin...
12	ADN1 TV	http://www.adn...	National	Newsroom	Entertainment	Television	https://www.facebook.co...	
13	ADNTV News	https://adntv.tv/	National	Newsroom	General News	Online	https://www.facebook.co...	https://www.linkedin...
14	Aeipath Studios	https://www.ae...	National	Production	Entertainment	Online	https://www.facebook.co...	https://www.linkedin...
15	AFCA (African Fact-checkin...	https://factche...	Continental	Investigative	General News	Online		
16	Afoto Films	https://afotofil...	National	Production	Entertainment	Online		https://www.linkedin...
17	Africa Decoded	http://africade...	National	Production		Online		
18	Africa Eco News	https://africaec...	Continental	Newsroom	Environment	Online	https://www.facebook.co...	
19	Africa Finest TV		Municipal					
20	Africa InSight Communicati...	http://www.afri...	National	Newsroom	General News	Online	https://www.facebook.co...	https://www.linkedin...

A screenshot of a sample on MediaData (CfA using Airtable)

5.2 Media monitoring

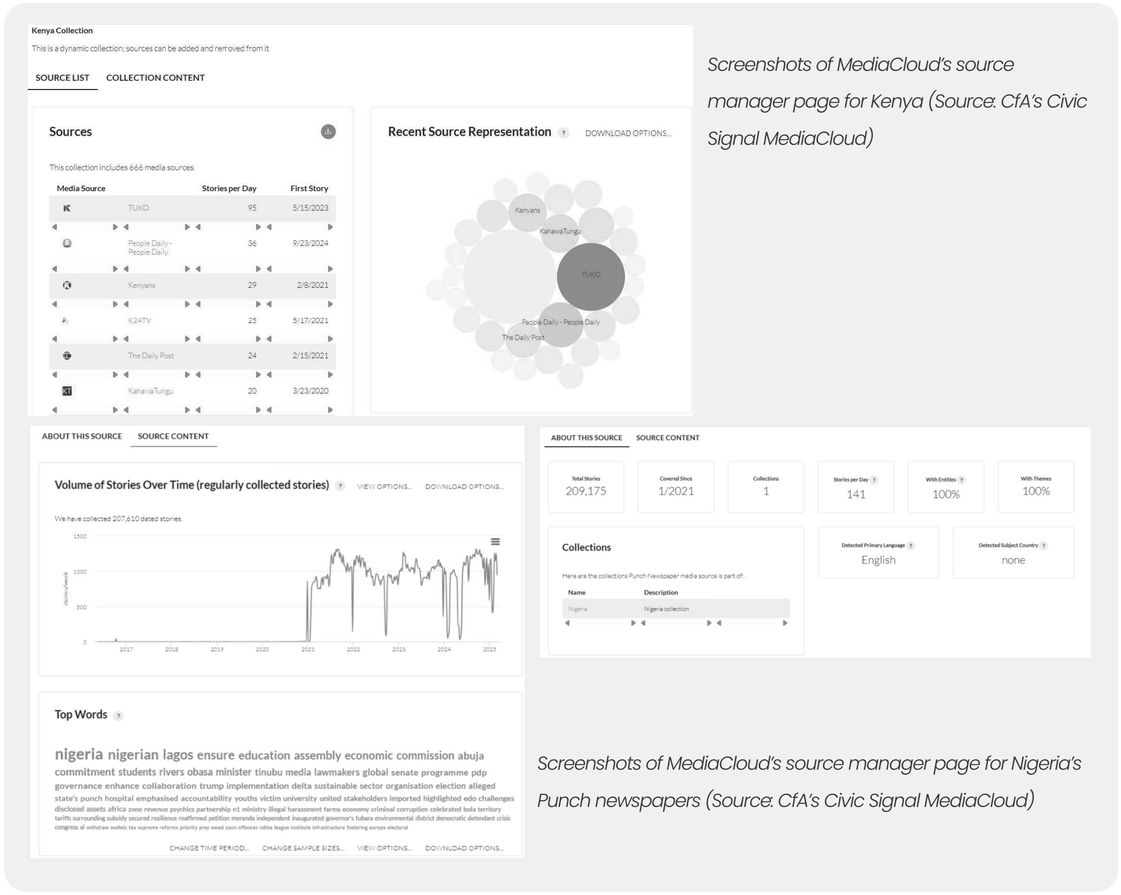
CivicSignal MediaCloud is an open-source, open-data platform which uses natural language processing (NLP) technology to gather and analyse digital news media content across Africa.
It is adapted from Media Cloud, which was developed at Harvard University's Berkman Klein Center and the Massachusetts Institute of Technology, and now belongs to the Media Ecosystems Analysis Group.
MediaCloud enables users to store, retrieve, visualise, and analyse news stories collected through continuously updated feeds.

CivicSignal's MediaCloud has two main parts: source manager and explorer.

a. MediaCloud source manager

This helps users organise and track media sources. A 'source' refers to any organisation – such as a fact-checking group, media lab, or newsroom – that publishes digital news. These sources are grouped into 56 country-specific collections. Users can explore collections to view daily story volumes, data collection timelines, and sources and export source information as needed.

A source page provides users with insights, including total story count, word clouds from term frequency in samples, a timeline of story collection, and geographic coverage based on extracted location data. This detailed view helps researchers analyse content patterns and coverage focus.



Screenshots of MediaCloud's source manager page for Kenya (Source: CfA's Civic Signal MediaCloud)

Screenshots of MediaCloud's source manager page for Nigeria's Punch newspapers (Source: CfA's Civic Signal MediaCloud)

b. CivicSignal MediaCloud explorer

Explorer allows users to search and analyse how digital news media covers specific topics. It provides insights through three dimensions:

- **Attention**: Tracks coverage volume by measuring the number of captured stories.
- **Language**: Generates word clouds based on term frequency analysis.
- **Entities**: Identifies the most mentioned people and organisations.

Step-by-step guidelines on how to use MediaCloud explorer

Initial setup

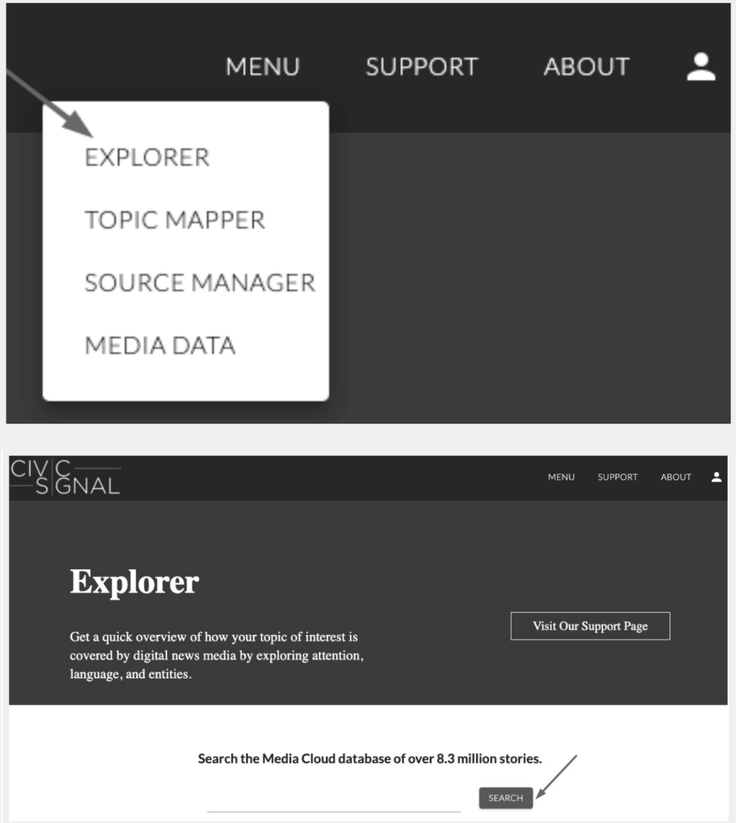
Visitors must log in or register before using explorer.

To start a search, go to **Menu** >> **Explorer** and **click SEARCH** to access the query interface.

Writing effective queries on MediaCloud

CivicSignal MediaCloud supports Boolean operators to refine searches:

AND: Requires both terms to appear (narrows results).
OR: Includes results with either term (broadens results).
NOT: Excludes specific terms.
Parentheses (): Group terms to control query logic.



Screenshot of how to access the MediaCloud explorer (Source: CfA's CivicSignal MediaCloud)

Put queries under ‘Enter search terms’. Here are some examples:

- election OR vote OR ballot – Finds stories with any of these terms.
- election AND misinformation – Finds stories mentioning both terms.
- (election AND misinformation) NOT (‘social media’ OR ‘Facebook’) - Excludes stories mentioning social media or Facebook.

Selecting a media source

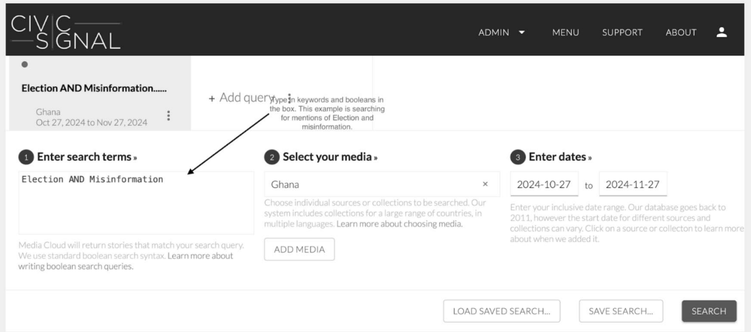
MediaCloud queries can target multiple collections, a single collection, or a specific media source. To set the search scope, click ‘**Add media**’ in the ‘**Select your media**’ section.

You can select media in two ways:

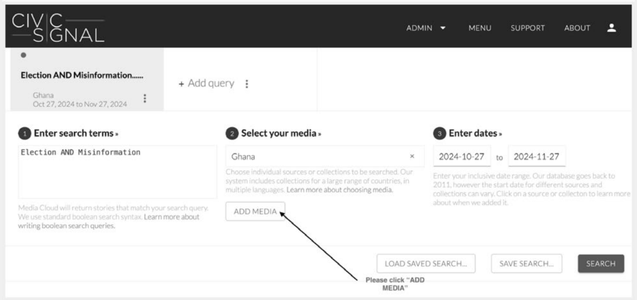
a. **Search sources and collections**

Click ‘**Search Sources & Collections**’.
Type the collection name (e.g., ‘Ghana’) in the search bar.
Click ‘**SEARCH**’.
Click the ‘+’ sign to select the collection.
Click ‘**OK**’.

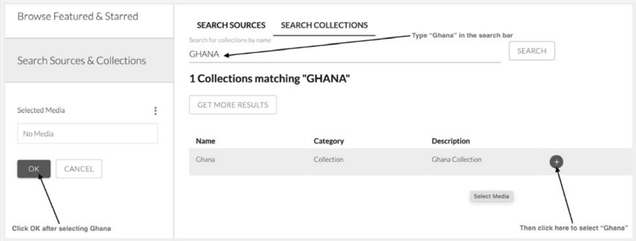
To search within a specific source, select ‘***SEARCH SOURCES***’ instead of ‘***SEARCH COLLECTIONS***’ and enter the source name.



Screenshot showing the process of entering search queries in the MediaCloud explorer (Source: CfA’s CivicSignal MediaCloud)



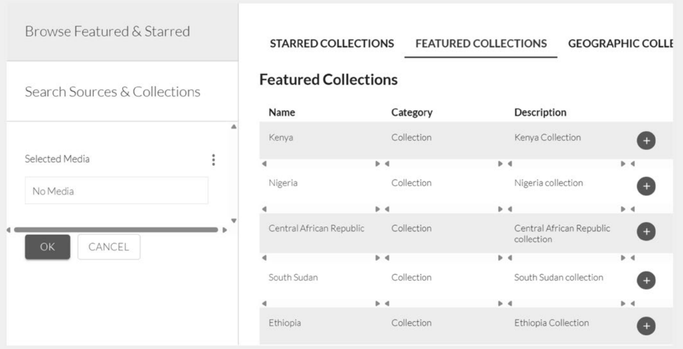
Screenshot illustrating the process of selecting media in the MediaCloud explorer (Source: CfA’s CivicSignal MediaCloud)



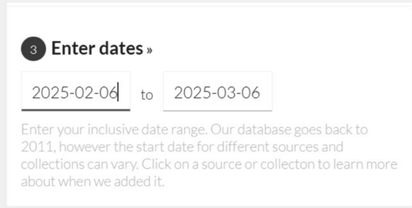
Screenshot illustrating the process of searching ‘Sources & Collections’ on MediaCloud (Source: CfA’s CivicSignal MediaCloud)

b. Using ‘Browse Featured & Starred’
Navigate to ‘Browse Featured & Starred’.
Locate and select the desired source from the list.
Specify the period of interest
Enter dates in YYYY-MM-DD format.

Once your query is set, click the ‘SEARCH’ button to search the database.



Screenshot illustrating the process of browsing featured collections on on MediaCloud (Source: CfA’s CivicSignal MediaCloud)



A screenshot illustrating the process of selecting the date range for the query on on MediaCloud (Source: CfA’s MediaCloud)

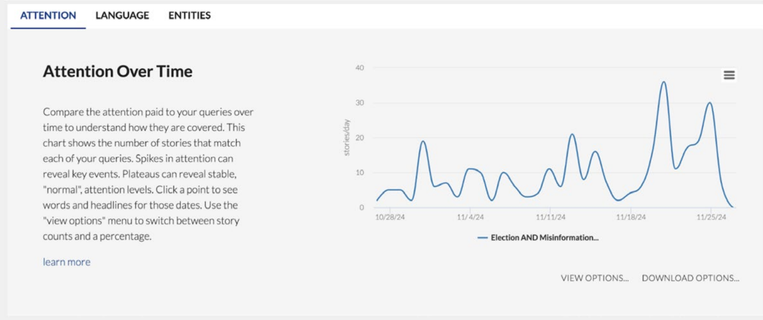
Query result analysis

MediaCloud provides three primary analysis tools, accessible through dedicated tabs: Attention, language, and entities.

a. Attention tab

The attention over time graph pinpoints peaks and declines in keyword mentions. Users can toggle between viewing the number or percentage of stories that include specific keywords relative to overall news coverage, depending on their search needs.

MediaCloud allows users to download search query results, including story URLs, in CSV format for further analysis.



A screenshot of MediaCloud’s attention tab (Source: CfA’s MediaCloud)

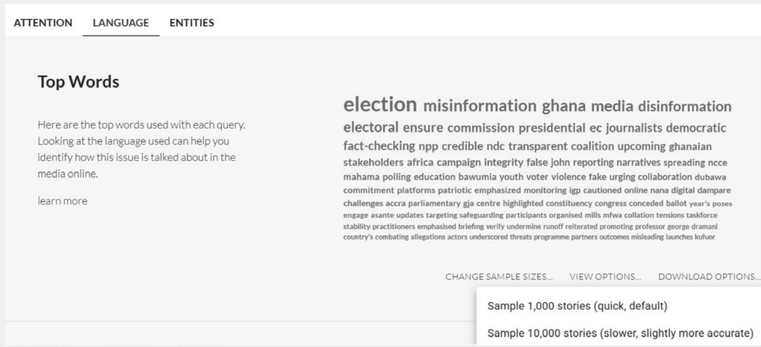
b. Language tab

MediaCloud visualises language patterns in search results using word clouds under the '**LANGUAGE**' tab. These visualisations highlight the most frequently mentioned terms in articles matching your query.

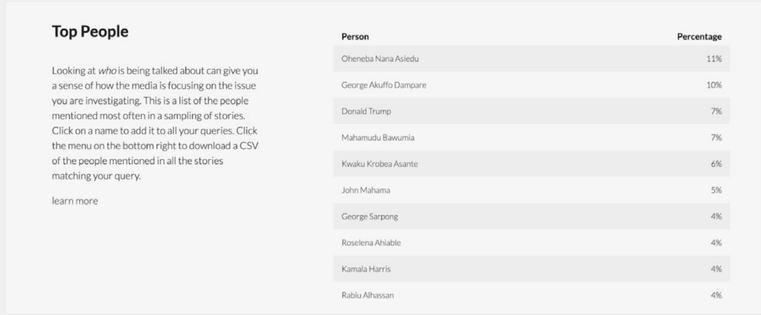
In the '**Top Words**' word cloud, text size reflects frequency – larger words appear more often. The system generates these visualisations from a representative sample of 1,000 stories by default, but users can expand this to 10,000 for a more comprehensive analysis, though this requires additional processing time.

c. CivicSignal MediaCloud highlights the top people and organisations mentioned in stories under the '**ENTITY**' tab.

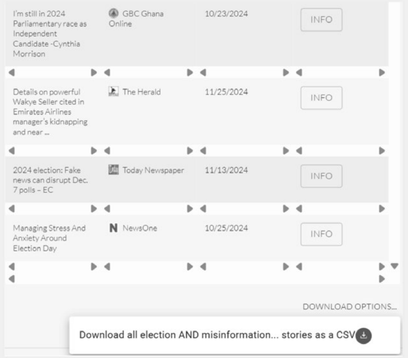
MediaCloud uses named entity recognition algorithms to extract and rank people and organisations mentioned in articles based on their frequency of appearance. The system calculates the percentage of articles containing each entity, helping researchers identify the most prominent actors in a given topic. Geographic coverage analysis maps the spatial distribution of places referenced in stories, revealing a publication's reporting focus beyond its home country. For instance, Nigerian outlet The Punch reports not only on domestic events but also on international news from locations such as the US. However, geographic tags appear only when the system successfully extracts and recognises location data.



A screenshot illustrating how to change the sample size for top words (Source: CfA's MediaCloud)



A screenshot showing the 'Top People' section under the MediaCloud 'ENTITY' tab (Source: CfA's MediaCloud)



A screenshot showing how to download query results from MediaCloud (Source: CfA's MediaCloud)

5.2.1 TrollTracker watchlists

A troll tracker watchlist is a structured database of entities, PIPs, political parties, and their associated social media accounts. The watchlist is used to track online conversations, detect emerging narratives, and set up early warning systems for coordinated influence campaigns, disinformation, or hate speech.

This guide provides a step-by-step process for developing a watchlist for social media investigations, integrating it into monitoring tools, and setting up an early warning system for a proactive response.

Define focus areas:

Develop and map a watchlist to track disinformation, hate speech, and political narratives across actors and entities.

Create a structured database:

Use tools such as Airtable, Excel, or Google Sheets to establish a database with relevant columns for actor categorisation and social media accounts.

Identify PIPs

Map out elected leaders, election candidates, and opposition figures who are influencing political discourse.

Map political parties and entities:

Document the political landscape, including coalitions and alliances, political groups, political parties, and youth leagues.

Identify political activists:

Identify accounts that are highly vocal and actively shaping political discussions online.

Categorise entities by political affiliations:

Classify each identified actor or entity as: neutral, independent, opposition-affiliated, or pro-government.

Conduct social media mapping:

For each mapped entity, document its presence on major social media platforms such as Facebook, Instagram, Telegram, TikTok, X, and YouTube.

Validate social media profiles:

Verify the authenticity and current status of mapped social media accounts, ensuring the accounts are still active and have not undergone significant profile changes, for example shifting from a personal account to a political campaign page.

Integrate social media profiles into monitoring tools:

Upload the verified list of social media accounts into social media intelligence (SOCMINT) tools to track conversations and narratives.

Configure keyword-based alerts:

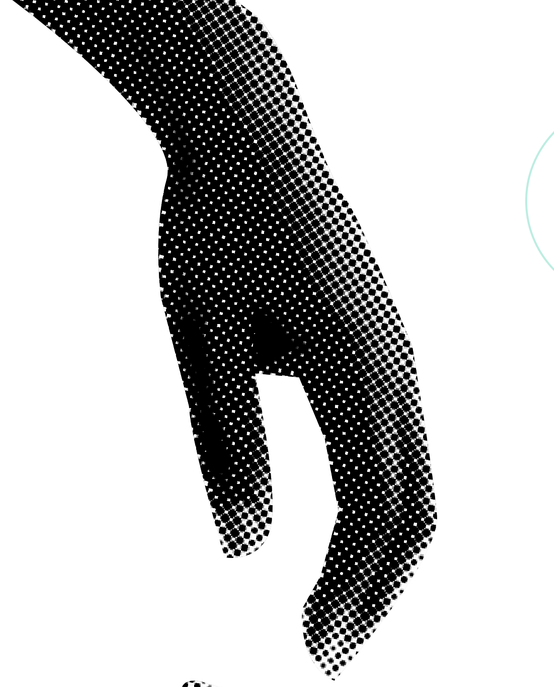
Set up real-time alerts and notifications to detect high-risk content that is originating from the troll tracker list, such as inciteful election-related posts, hate speech, and inflammatory rhetoric.

By following these steps, the watchlist will provide early warnings and allow the proactive monitoring of political discourse and online threats.

5.3 Human Intelligence (HUMINT)

Human intelligence refers to information gathered from human sources, typically through direct interaction, such as interviews, conversations, or informal interactions.

This type of intelligence relies on human insights and perceptions to provide clues, contexts, and actionable information that may not be available through more technical means such as satellite imagery or social media investigations. It plays a crucial role in understanding motivations and trends within a target population.



5.3.1 Media Sentinels

Media Sentinels are experienced journalists with social media monitoring expertise and deep community knowledge. They track information manipulation and violent extremism on dark socials and other platforms, provide early warnings for offline events that may spark misinformation, offer local context for research and investigations, and contribute insights to analysis reports.

Working with Media Sentinels

Step 1: Adopt an audience-first approach

a. Use partnerships with social media platforms to analyse target audiences based on specific countries or regions.
b. Conduct behavioural, psychographic, and sentiment analysis to inform monitoring strategies.

Step 2: Identify and vet sentinels

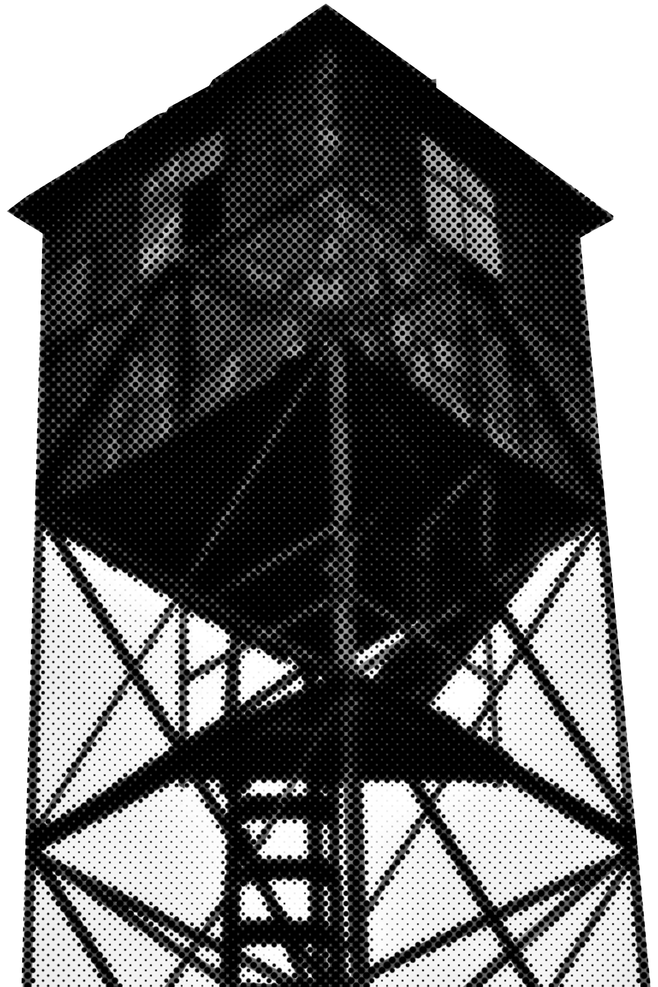
a. Recruit individuals with strong community ties and journalistic integrity.
b. Ensure neutrality by reviewing their political affiliations and professional backgrounds.
c. Select sentinels based on access to specific communities, alignment with monitoring goals, and credibility.

Step 3: Develop monitoring frameworks

a. Create adaptable frameworks that support accurate information tracking.
b. Align sentinel expertise with overarching programme objectives.
c. Define platforms, resources, and themes for effective monitoring.

Step 4: Provide guidance and support

a. Maintain open communication through relationship managers.
b. Offer training, resources, and real-time support on sensitive topics.
c. Ensure alignment with programme priorities while respecting sentinel expertise.



Step 5: Share insights

a. Sentinels contribute verified, community-grounded insights.
b. Provide early detection of misinformation before it spreads to mainstream platforms.
c. Enable data-driven interventions based on authentic, closed-network perspectives.

Step 6: Assess programme impact

a. Conduct regular evaluations of sentinel effectiveness.
b. Analyse trends, emerging threats, and strategy effectiveness.
c. Adjust monitoring approaches to enhance impact and reach.

5.3.2 Tiplines

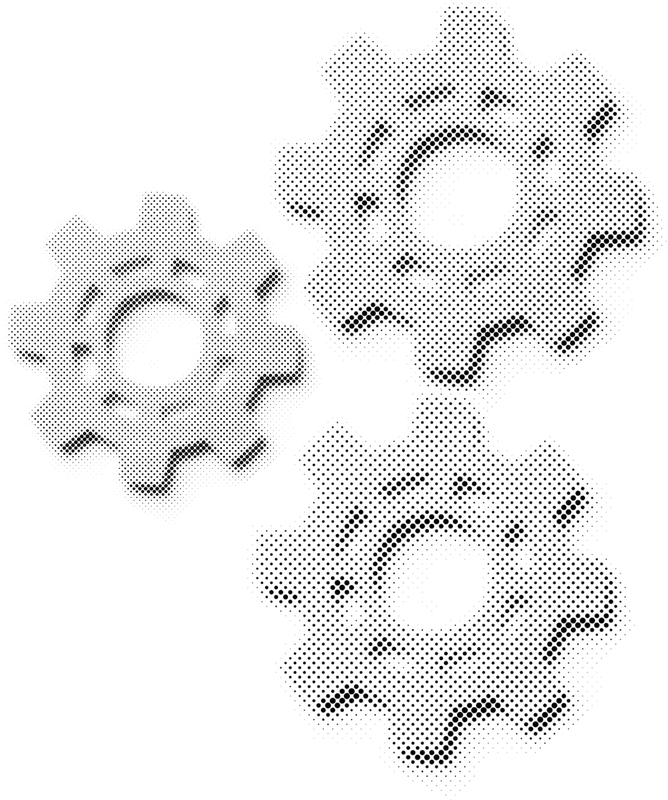
A tipline is a channel for crowdsourcing misinformation leads via chatbots, email, hotlines, SMS, social media, or websites. Organisations can establish these tiplines on various platforms, including email, phone calls, Slack, SMS, Telegram, Web, WhatsApp, or via an API.

CfA runs a centralised tipline on Check, consolidating reports from Facebook, X, and PesaCheck's WhatsApp. It offers instant 'claim matching', immediately comparing new information with a database of existing fact-checked claims, reducing verification time.

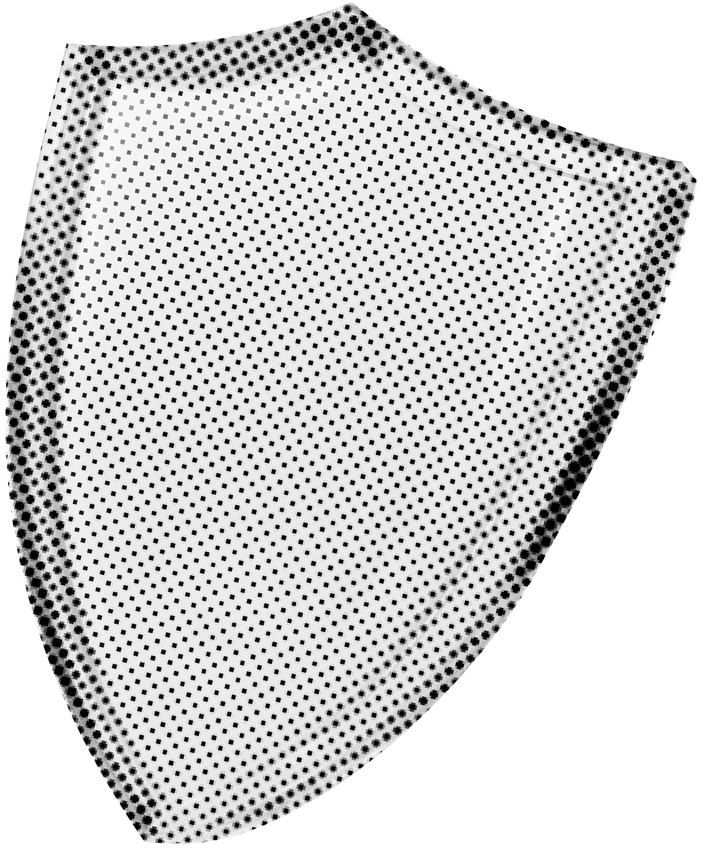
Tiplines allow fact-checkers to track and dispel misinformation on closed platforms, while also monitoring trending topics and online discourse. It encourages readers to submit claims, but also to share the debunked information with their groups. During election times, a tipline enables fact-checkers to work in real time to curb the spread of incorrect information.

Step-by-step guide to setting up a tipline

1. **Define the purpose:** Determine whether the tipline will handle general fact-checking or focus on specific topics such as elections, climate, or gender misinformation.
2. **Establish the scope:** Set clear criteria for fact-checkable content, including content type, geographic reach, and topic.
3. **Choose a communication channel:** Select a platform for tip submissions (email, online forms, phone, SMS, social media, or web-based apps).
4. **Set privacy and terms of use:** Ensure the security of user data with a clear privacy policy and encrypted communication where necessary.
5. **Simplify submission:** Provide clear instructions and make the tip submission process user-friendly.
6. **Build a review team:** Assemble a team to verify and categorise tips based on predefined fact-checking criteria.
7. **Establish a follow-up system:** Respond to every submission with a fact-check or relevant feedback, automating standard replies where possible.
8. **Promote the tipline:** Make it visible on social media, websites, and other platforms to increase engagement.
9. **Monitor performance:** Analyse trends, collect user feedback, and track submissions to refine the system.



5.4 Human rights defenders (minorities, refugees/migrants, etc.)



Human rights defenders (HRDs) and journalists play a crucial role in ensuring free and fair elections by promoting credible information and countering disinformation.

Their work includes monitoring electoral processes, exposing human rights violations, and fostering transparency. Effective collaboration between civil society, investigative bodies, and media organisations strengthens their impact.

How to work together effectively

- **Build a stakeholder network:** Connect with journalists covering elections and HRDs focused on electoral justice to form a coalition.
- **Establish clear goals:** Align on shared objectives such as ensuring accurate information, transparency, and voter rights while maintaining non-partisanship.
- **Training and capacity building:** Conduct workshops to equip communities, HRDs, and journalists with skills to detect and combat disinformation.
- **Information and resource sharing:** Use shared online platforms and communication channels to verify and distribute credible election-related information.
- **Election monitoring:** Track and analyse disinformation trends, leveraging technology to identify misinformation sources and report irregularities.
- **Rapid disinformation response:** Deploy a team to issue timely corrections and counter false narratives across media channels.
- **Advocacy and public engagement:** Push for policies protecting civil society, free speech, and media literacy while engaging communities in outreach initiatives.
- **Safety and security measures:** Provide legal support, safety training, and secure communication tools to HRDs and journalists facing threats or harassment.
- **Evaluate and adapt strategies:** Assess the effectiveness of efforts, refine approaches based on feedback, and adjust tactics as needed.
- **Post-election reflection:** Conduct debrief sessions to document challenges, lessons, and successes, informing future election interventions.

5.5 Behavioural analysis

Addressing threats from malign actors usually requires understanding the tactics they are using to disrupt the information ecosystem, in other words, analysing their behaviours.

5.5.1 Killchains + phase-based analysis

A killchain is a structured model breaking threat actor operations into phases, such as preparation, execution, and post-execution, to identify intervention points and patterns.

Modern analyses of online influence operations emphasise three core phases: Preparation, execution and post-execution.

How to use a killchain model:

a. Map campaigns to hybrid phases

i. The table beside outlines the core phases of influence operations:

b. Identify recurring tactics

i. Preparation: Use of forged personas (e.g., DISARM TTP: T0011 – fake accounts).
ii. Execution: Algorithmic gaming (e.g., hashtag hijacking: TTP: T0049 – Flood information space).
iii. Post-execution: Repurposing old content (e.g., TTP: T0098 – Recycled narratives).

c. Phase-specific countermeasures

i. Preparation: Preemptive exposure of fake accounts (Blue TTP: T0011.001).
ii. Post-execution: Partner with platforms to archive deleted content for forensic analysis.

Phase	Question	DISARM phase
Preparation	Audience segmentation and asset creation	Planning/preparation
Execution	Cross-platform amplification	Execution
Post-execution	Metrics tracking and tactic adaptation	Evaluation

DISARM killchain

The DISARM killchain focuses on the lifecycle of disinformation campaigns, structured into four phases: planning, preparation, execution, and evaluation. Unlike traditional killchains, DISARM emphasises narrative development and iterative refinement. For example:

i. **Planning:** Adversaries define goals (e.g., election interference).
ii. **Preparation:** Create fake accounts and content libraries.
iii. **Execution:** Coordinate posting across platforms.
iv. **Evaluation:** Measure engagement and adjust tactics.

Here are the steps for using the killchain.

i. **Phase mapping:** Use DISARM's taxonomy to assign campaign actions to phases.
ii. **Analyse interdependencies:** Example: Preparation-phase bot farms enable execution-phase amplification.
iii. **Counterphase actions:** Deploy blue-team tactics:
    i. Planning: Expose adversary goals via intelligence sharing.
    ii. Execution: Slow down or limit bot activity

Meta killchain

The Meta killchain integrates various killchain models (e.g., Cyber Kill Chain and DISARM) to analyse multi-phase, cross-platform campaigns. It is especially useful for complex campaigns such as hybrid warfare, where cyberattacks and disinformation work together.

By layering different models, the Meta killchain uncovers how adversarial actions in one domain (e.g., hacking) can trigger or support disinformation in another, allowing for a more comprehensive understanding and response to such threats.

Here are the steps for using the Meta killchain.

i. **Composite modelling:** Overlay DISARM phases with traditional killchain stages.

ii. **Cross-domain analysis:**
   i. Example: A data breach (Cyber Kill Chain's exploitation) fuels disinformation narratives (DISARM's execution).

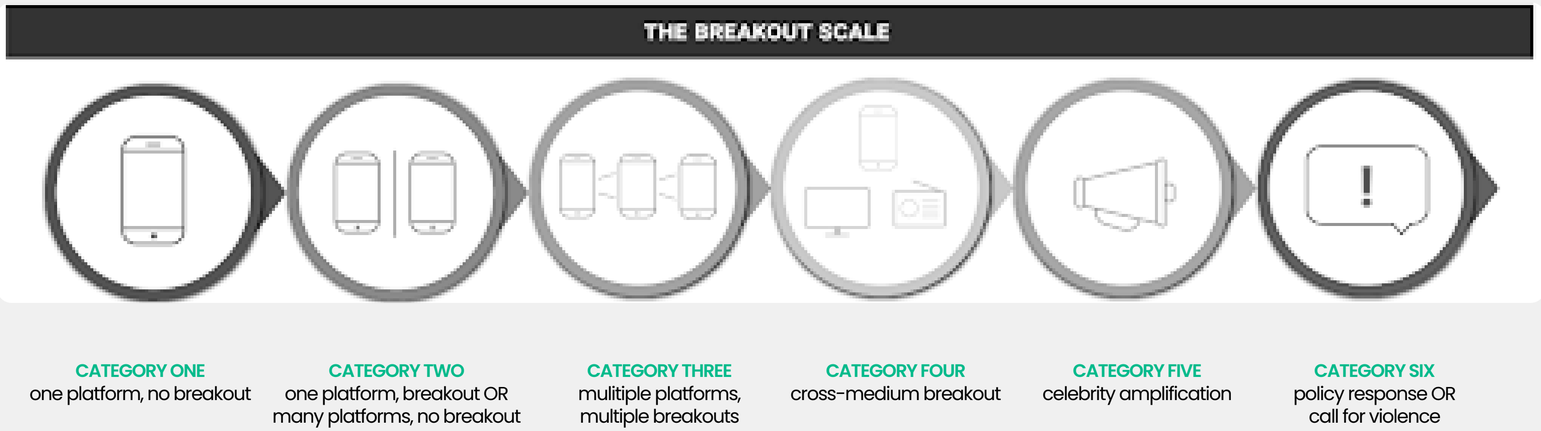
iii. **Holistic mitigation:** Develop cross-functional responses (e.g., using cybersecurity teams and fact-checkers).

Phase	DISARM activity	DISARM phase	Question
Preparation	Bot farm creation	Weaponization (malware tools)	Take down fake accounts + block IP addresses
Execution	Amplify forged documents	Delivery (phishing emails)	Email filters + public alerts

5.5.2 Barometers

Breakout scale

The breakout scale provides a framework for assessing the severity of influence operations (IOs) based on their reach and impact. It helps platforms determine the appropriate response by categorising harmful activities across six levels. Here is a step-by-step guide on how to analyse influence operations using the breakout scale:



The Breakout Scale shows a way of measuring the impact of an influence operation without having to measure changes in the behaviour of those targeted (Source: CfA via Brookings Institution)

Steps to analyse influence operations using the breakout scale:

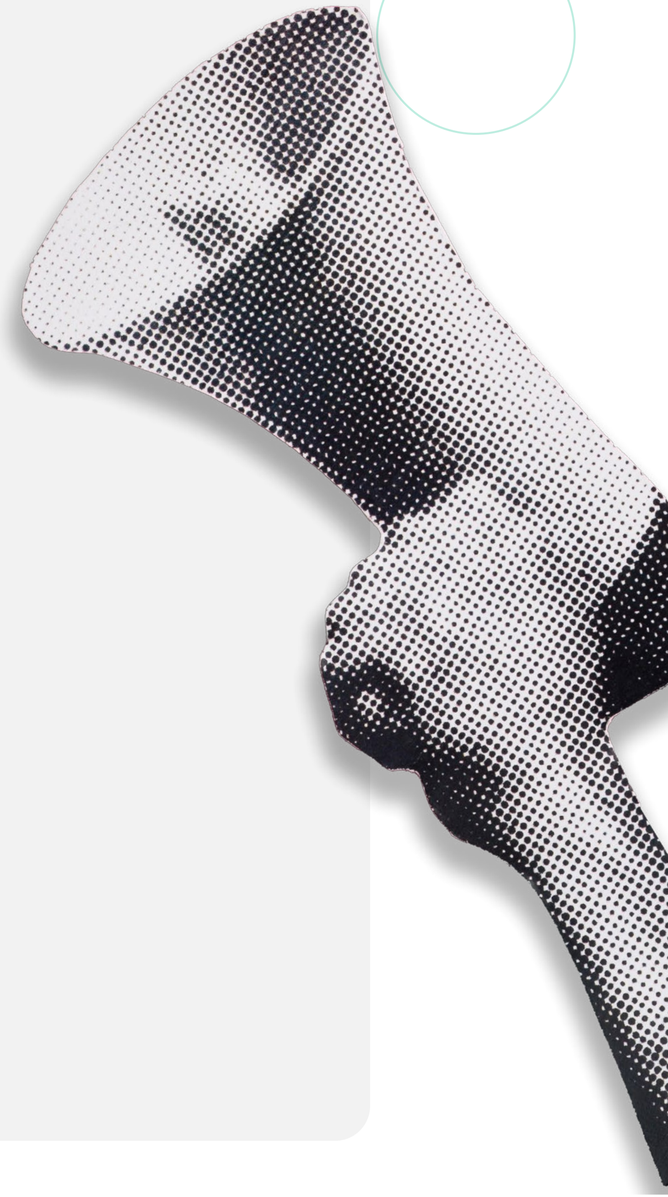
i. **Gather information**: Collect comprehensive information about the influence operation, including the platforms it operates on, the content it disseminates (such as images, messages, and videos), its target audience and communities, amplification efforts such as media coverage and shares, and the observed effects or reactions, such as policy changes or public discussions.

ii. **Determine platform scope**: Identify the platforms where the influence operation operates, determining whether it is confined to a single platform such as Facebook or X or spreading across multiple channels such as blogs, news websites, and social media.

iii. **Assess community reach**: Assess the reach of the influence operation by determining whether it targets a specific community, such as an interest group or local area, or extends to multiple, diverse communities.

iv. **Categorise based on platform and community**: Classify the influence operation based on platform and community reach:
   - **Category one**: Limited to a single platform and a single community.
   - **Category two**: Either a single platform reaching multiple communities or multiple platforms targeting a single community.
   - **Category three**: Spanning multiple platforms and engaging multiple communities.

v. **Evaluate amplification**: Assess amplification beyond social media:
   - Mainstream media pickup: Has the influence operation been covered by newspapers, radio, or TV?
   - High-Profile endorsement: Have influential figures, such as celebrities or politicians, amplified the influence operation?



Steps to analyse influence operations using the breakout scale:

vi. **Adjust the category accordingly**:

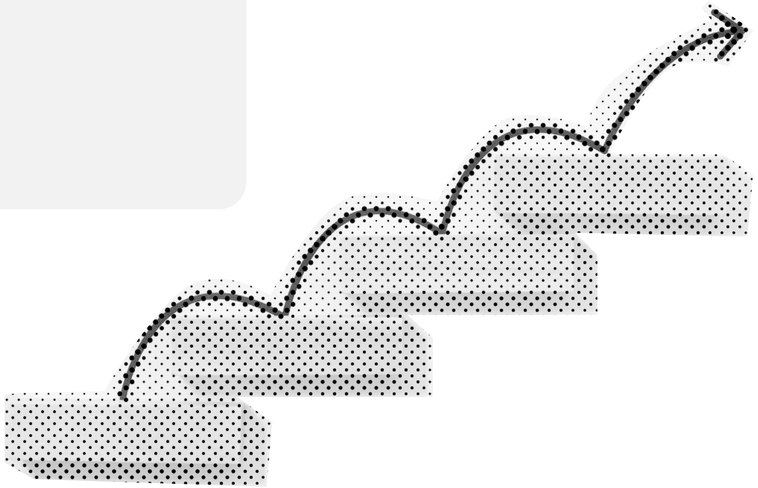
- **Category four**: Mainstream media amplification.
- **Category five**: High-profile individual amplification.

vii. **Assess impact and actions**: Determine whether the influence operation has had tangible effects, such as influencing policy through new laws or regulations, prompting real-world actions like protests or boycotts, or including calls for violence or incitement.

viii. **Assign the highest category if the answer is yes to any of the above**:

- **Category six:** Calls for violence, concrete action, or policy response.

ix. **Document and track**: Document the analysis by noting the assigned category and supporting evidence, and continuously track the IO for any changes in amplification, impact, or reach, updating the categorisation as necessary.

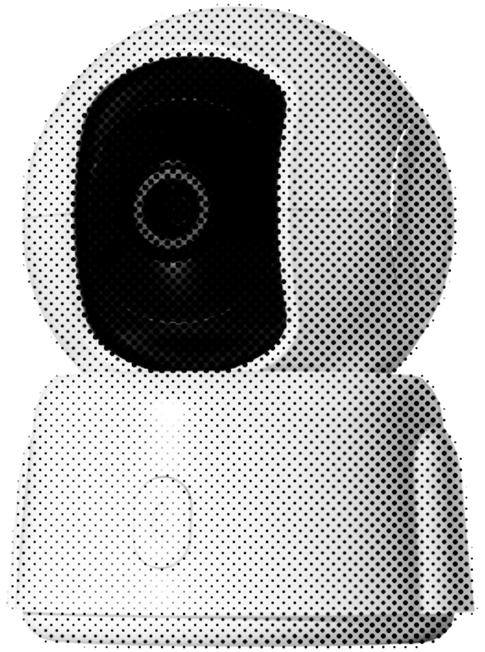


6.

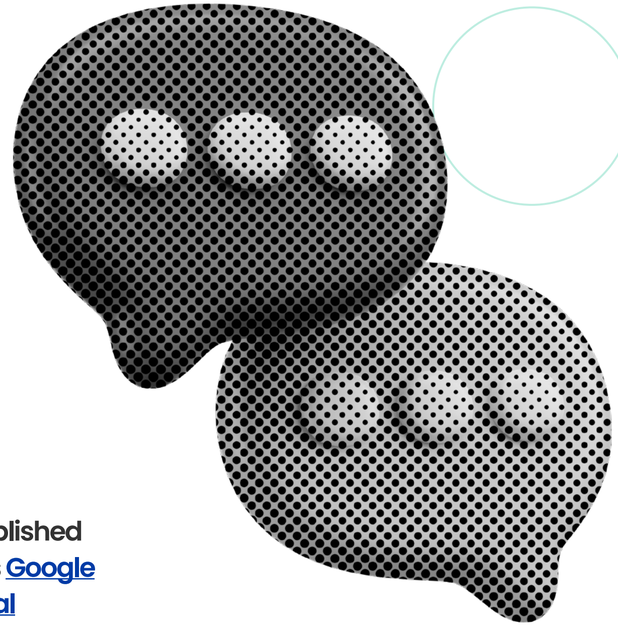
Detecting and analysing online threats

6. Detecting and analysing online threats

This section provides resources and step-by-step methods for monitoring activities on open and dark social media. It covers how extremist groups evade detection, techniques for investigating dark web communication, and OSINT tools for verifying disinformation in images, videos, and audio. It also explores the importance of mapping social media networks.



6.1 Social media intelligence (SOCMINT)



To map social media platforms and their role in political mis-/disinformation, use these data sources:

Statista
Provides reports on social media use, platform rankings, and trends in political content.

Similarweb
Offers traffic insights to identify popular platforms and audience behaviour.

We Are Social/Hootsuite Reports
Provide annual reports on global social media use, identifying platforms for political content.

Research studies published on platforms such as Google Scholar and the Social Science Research Network (SSRN)
Include academic journals and reports on social media's impact on political discourse and mis-/disinformation.

Appfigures
Tracks mobile app downloads and rankings to highlight popular platforms.

Statcounter
Analyses platform market share by country to spot trends in political discussions.

For effective monitoring, categorise platforms as '**dark**' (e.g., encrypted) or '**open**' (e.g., public-facing). Combining these sources with independent research gives a comprehensive view of social media's role in political misinformation.

6.1.1 Open social

To monitor election disinformation on open social media platforms, follow these steps:

Use social monitoring tools:

Tools such as Brandwatch and Meltwater use a ‘search’ principle to track keywords, names, or other relevant information. These platforms help analyse narratives or specific entities related to disinformation.

Iterate queries:

Disinformation trends evolve quickly, so continuously update your queries with new terms. Multiple queries may be required to track different narratives, as search terms grow longer and harder to manage.

Set up alerts:

Configure alerts based on mention frequency within a set timeframe. Alerts help pinpoint peaks in discussions, highlighting events that may require closer investigation.

Streamline the process:

Use alerts and refined queries to focus on high-priority events and topics, ensuring you stay ahead of disinformation trends.

6.1.2 Dark social

To monitor dark social platforms such as Signal, Telegram, and WhatsApp for election-related disinformation, follow these steps:

Identify groups or channels:

Start by searching social media platforms such as Facebook or X for links to WhatsApp or Telegram groups. Use terms such as 'chat.whatsapp.com' or specific keywords related to political groups or parties to narrow down your search.

Monitor public groups or channels:

While private messages on encrypted platforms are inaccessible, public groups or channels can be monitored. WhatsApp channels, in particular, are public-facing and can be easily joined or subscribed to. These channels, though new, can be important for political campaigns.

Use specific search terms:

To refine your search, combine the broad 'chat.whatsapp.com' term with relevant keywords tied to political parties or movements. This helps in finding groups with specific election-related content.

Monitor and track activity:

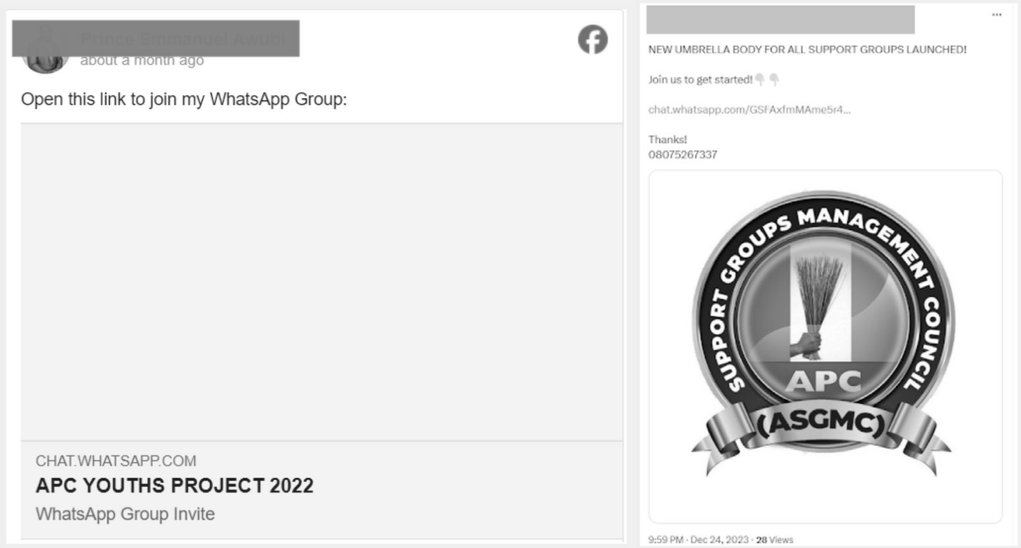
Once you have access to the groups or channels, track any discussions, shared media, or false narratives related to the election. Stay alert to narratives, trends, or spikes in activity that may need deeper investigation.

Real-life example

To identify WhatsApp groups related to political parties in Nigeria, such as the All Progressives Congress (APC), the Labour Party, and the Peoples Democratic Party (PDP), you can refine your search queries as follows:

1. **Search by party acronym and candidate names:** Modify your search to include the political party's acronym or the names of party candidates. For example:
   a. On Facebook and X, use the query 'chat.whatsapp.com, APC' to find posts with WhatsApp group links mentioning the APC acronym.
   b. For the Labour Party, use 'chat.whatsapp.com, LP'.
   c. For the PDP, use 'chat.whatsapp.com, PDP'.

2. **Refine search with candidate names:** To narrow it down further, include the names of party candidates. For example:
   a. For the APC, you can search for 'chat.whatsapp.com, Bola Tinubu' or 'chat.whatsapp.com, APC, Tinubu'.
   b. For the Labour Party, search for 'chat.whatsapp.com, Peter Obi' or 'chat.whatsapp.com, LP, Peter Obi'.
   c. For the PDP, search for 'chat.whatsapp.com, Atiku Abubakar' or 'chat.whatsapp.com, PDP, Atiku'.



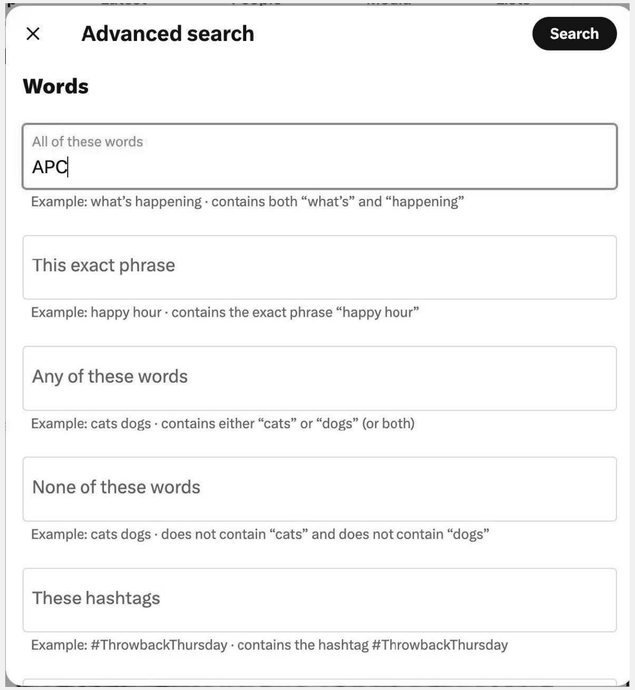
Screenshots showing a Facebook and an X posts sharing links to WhatsApp groups affiliated with the APC (Source: CfA via Facebook and X)

On X, the advanced search feature offers a user-friendly menu for searching posts by accounts, date of posting, keywords, and number of engagements.

Real-life example

The same query from the previous example can be searched on X using the advanced search feature in the following way:

We can specify an account to search for and filter unwanted words to remove false positives.



Screenshots illustrating the process of searching for WhatsApp groups related to political parties in Nigeria on X (Source: CfA via X)

Although the interactive advanced search menu is easy to use, it is currently not available through the X mobile app, so knowledge of some query filters will be useful to investigators. Below are some query filters and their applications.

Query filters	Usage
filter:replies	Filter for only reply posts
filter:verified (posts from accounts with a blue check mark)	Filter for posts made by accounts with blue check marks (X verified accounts)
filter:media	Filter for posts with media (video, pictures, etc.)
filter:news	Filter for posts from accounts identified as media pages on X

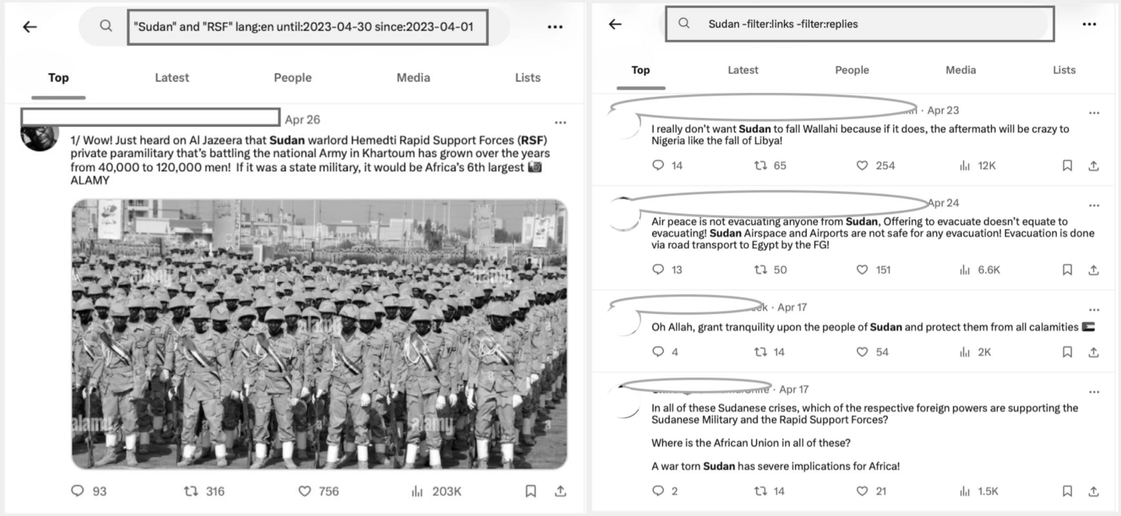
Query filters	Usage
from:account	Filter for posts by a particular account
to:account	Filter for posts mentioning an account or made in response to the specified account
since: (yyyy-mm-dd)	Filter to specify the starting date for search results
until: (yyyy-mm-dd)	Filter for posts by a particular account
- (hyphen)	Putting a hyphen in front of a filter query negates it. For example, '-filter:media' returns results from posts that do not contain media.

Real-life example: Using X advanced search for Sudan mentions

The screenshots on the right shows sample queries with X advanced search. The first query filters for posts in English that mention ‘Sudan’ and ‘RSF’ between 01 and 30 April 2023.

We can also negate queries by putting a hyphen before the queries. In the example below, we search for X posts that mention ‘Sudan’ but filter out posts that have links or are replies.

The same search principle applies to Telegram channels, using the term ‘t.me’, which is found on all Telegram channel links.



Images sourced from X/ Twitter

When investigating WhatsApp and other private groups, investigators must:

Follow legal guidelines:

Adhere to privacy and data protection laws. Anonymise data and report incidents, not individuals. In cases of an imminent threat, involve law enforcement.

Respect privacy:

Report findings ethically, focusing on the content, not individuals, to protect privacy and avoid legal issues.

Join early and regularly:

Group links reset over time, so early and consistent access is important.

Maintain anonymity:

Use anonymous or secondary accounts to avoid detection. Be cautious about SIM card registration that may expose personal information.

Fediverse

The multiplatform communication paradigm (MCP) allows extremist groups to evade detection by decentralising their operations across multiple small platforms. These groups utilise:

- **Beacons:** Main platforms for disseminating extremist content.
- **Content aggregators:** Online platforms for storing and sharing extremist material.
- **File stores:** Platforms where content is uploaded for later distribution.

This decentralisation complicates deplatforming efforts, as no single authority controls these platforms. Monitoring and tracking activities on such platforms often requires trial and error.

A space where MCP thrives is the Fediverse, a decentralised network comprising platforms such as Mastodon (an alternative to X), PeerTube (an alternative to YouTube), and Pixelfed (an alternative to Instagram). Each instance within the Fediverse is independently operated but communicates with others, offering privacy, freedom, and a means for extremists to evade moderation.

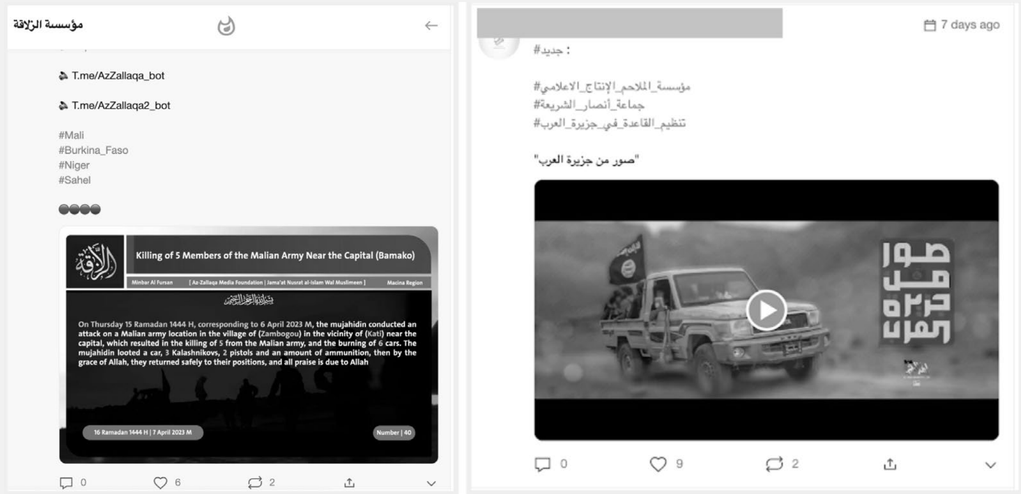
To effectively investigate extremist activities in this environment, methodologies must be adaptive, targeting individual instances, monitoring cross-platform content flows, and leveraging collaborative efforts across security and monitoring groups.

Chirpwire

Chirpwire is a social media platform similar to X. extremist groups, including Jama'at Nusrat ul-Islam wa al-Muslimin (JNIM) in the Sahel region, primarily use it to amplify their messaging.

CfA identified 245 posts on Chirpwire using the hashtag '#AQIM' [Al-Qaeda in the Islamic Maghreb] and tags such as 'Jamaat Supporting Islam and Muslims', which amplify and praise JNIM's activities. A sample of these posts can be found on the right.

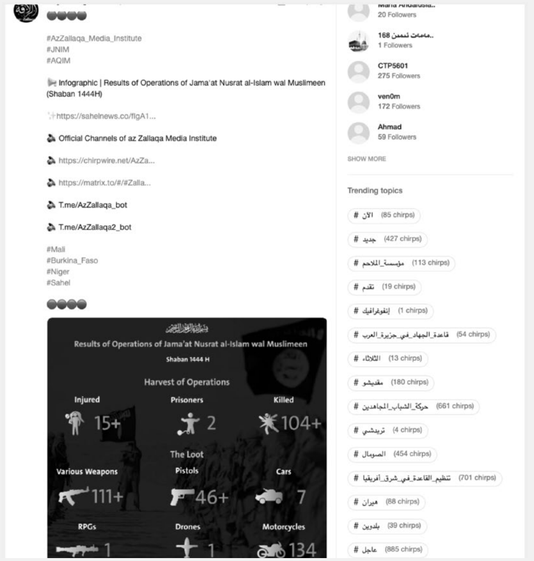
Furthermore, posts on the platform always redirect users to other platforms such as Matrix, Rocket Chat, and Telegram for more information. The redirection aims to create an ecosystem of accounts working together to amplify content and evade detection.



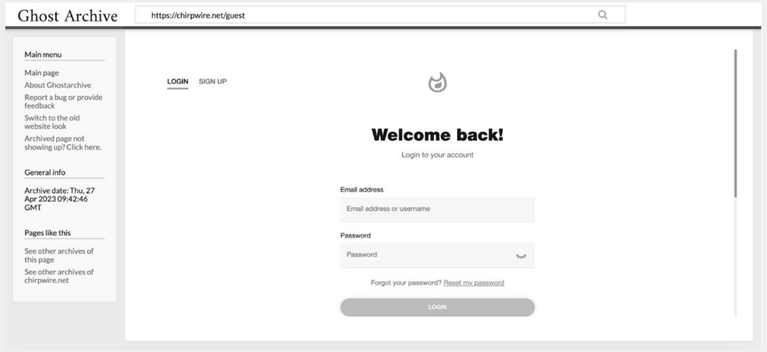
Sample screenshots from Chirpwire showing extremist content in Mali (left) and the Arabian Peninsula (right) (Source: CfA using Chirpwire) Translation of the right screenshot: [#new #Al-Malahem_media_production_establishment #Al Qaeda_in_the_Arab_Jazeera "Pictures from Arabia"]

Chirpwire offers comparable features to those of X, including the option to follow accounts, like posts, comment, and share content. However, the platform's ease of joining poses a significant risk. Chirpwire can be accessed through a basic search engine such as Google. In addition, individuals can easily register to Chirpwire using their existing social media accounts, such as Facebook, Google, Telegram, and X.

Furthermore, Chirpwire has implemented safety measures to safeguard the content posted on the platform, such that it cannot be archived using tools such as archive.today or ghost archive, as demonstrated below. This design feature ensures the privacy of posts made on the platform, such that digital archives such as Wayback Machine or other secondary platforms cannot reference the content.



A screengrab of a Chirpwire post containing links to other platforms such as Matrix and Telegram (Source: CfA using Chirpwire)



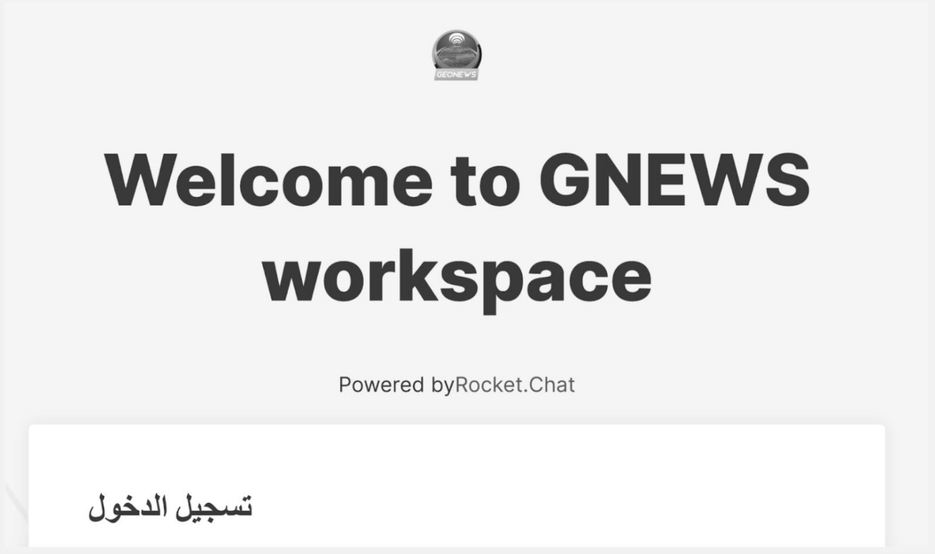
Screenshots from Ghost Archive showing failed attempts to archive Chirpwire content (Source:CfA using Archive Today and Ghost Archive)

Rocket Chat

Rocket Chat is a decentralised platform where users can host their own servers, where all communication data is stored and shared, allowing them to fully control it. This contrasts Facebook, Google, X, and other Web2-based social media platforms, which store user data on company-owned servers. This ensures the privacy of the content shared on the platform in the following ways:

1. On Rocket Chat, only the user of the server has access to the data and they can delete it permanently. This is unlike tech platforms such as Facebook, Instagram, and X, where deleted content may be retrieved if there is a need.
2. Large tech platforms use new technologies such as side-scanning to detect illicit content users store on centralised servers such as iCloud, including child and sexual abuse materials. However, data on decentralised servers is not subjected to such scans, which can be applied to to detect and combat extremist content.

Extremist groups in the Sahel use Rocket Chat because of the platform's versatility, unlike centralised platforms such as Facebook and X.

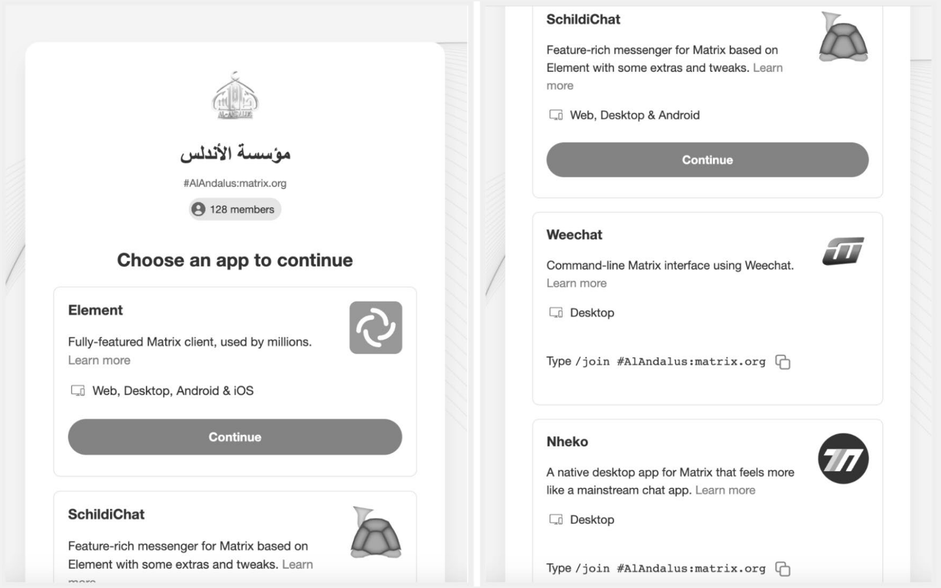


A screenshot of the sign-up page for GeoNews, an Al-Qaeda propaganda media outlet hosted on Rocket Chat, shared on Chirpwire (Source: CfA using Rocket Chat) [Translation of screengrab: Sign in]

Matrix

Matrix is a decentralised open communication platform used to send messages and make calls without requiring users to install a specific app or use the same application. It can be used on both desktop and mobile devices. The Salafi-Jihadi movement often recommends Matrix 'due to its utility, security, and anonymity'.

Al Andalus, a media outlet that disseminates propaganda for AQIM, shared links to Matrix on the Chirpwire platform.



Screengrabs showing Al Andalus using its Matrix platform to share links to various apps where users can access its content (Source: CfA using Matrix) [Translation of the username on the left screengrab: Andalus Foundation

How to investigate the dark web

Old-fashioned querying

The screenshots on the right shows sample queries with X advanced search. The first query filters for posts in English that mention 'Sudan' and 'RSF' between 01 and 30 April 2023.

For example, using lexicon terms identified an X account created on 15 August 2022, with 69 followers. The account promotes JNIM's media arm, Az-Zallaqa. The screenshots below show the account's systematic behaviour.

Many Fediverse platforms also provide search options that can be used to search for hashtags, keywords, and usernames.

Reverse image search

Reverse image search can be a powerful tool in uncovering networks of similarly structured yet inaccessible websites, particularly when traditional URL tracking methods fail. By uploading an image, such as a logo, infographic, or manipulated photograph, investigators can identify other sites using the same or similar visuals. This technique often reveals clusters of coordinated websites that share common design elements, host identical content, or promote the same narratives, despite lacking direct hyperlink connections.

Screenshots showing Az-Zallaqa's Chirpwire updates shared on X (left, centre, right (Source: CfA using X)

Real-life example: Uncovering the I'lam Foundation's network through reverse image search

A now-defunct website, I'lam Foundation, shares content promoting hate, radicalisation, toxic speech, and violent extremism. It is a global news outlet that publishes extremist content weekly, targeting Africa and the Middle East. The website has a South African domain and is web-hosted by Cloudflare, a US-based company specialising in security and performance for websites. Attempts to use the URL-based approach to track any content shared from the website to large mainstream social media platforms were unsuccessful.

However, reverse image search tracked online copies of the website's content. The search found websites with domains structured similar to that of I'lam Foundation's, but the sites failed to load any data and displayed an error message saying, 'This site can't be reached'. The websites include:

i. https://i3lam.blog/77758/
ii. https://i3l.today/lang/french/fr_naba/
iii. https://i3l.ltda/258486/
iv. https://i3l.today/260312/
v. https://i3l.website/lang/russian/ru_naba/

Another Cloudflare-hosted website, raud.wf, also republished the content, but it is currently non-functional.

Screengrabs showing the I'lam Foundation website sharing toxic content targeting Sahel and Central Africa (Source: CfA via the now-defunct I'lam Foundation website)

Web3

Web3 represents a paradigm shift in the evolution of the internet, prioritising decentralisation, user autonomy, and blockchain technology. While this development offers benefits such as enhanced privacy and control, it also presents new challenges for monitoring and regulating harmful content, particularly when extremist groups take advantage of these technologies.

Current most active Web3 tools extremist groups use:

i. **Matrix and Rocket Chat:** These decentralised messaging platforms allow for private, secure communication that is not controlled by any central authority. Extremist groups can use them to coordinate activities, share content, and engage in conversations without fear of censorship or surveillance.

ii. **InterPlanetary File System(IPFS):** IPFS is a distributed file storage system that allows users to store and retrieve content without reliance on a central server. This decentralised nature makes it difficult for authorities to block or censor content, providing extremist groups with a secure and anonymous way to distribute documents, propaganda, and videos.

iii. **Monero:** Monero is a privacy-focused cryptocurrency that enables transactions to be conducted without revealing the senders' or receivers' identities, offering anonymity in financial dealings. Extremist groups use Monero to raise funds and make transactions without detection, complicating efforts to track illegal financial activities.

iv. **Ethereum Name Service (ENS)/EthLinks:** Websites built on the Ethereum blockchain using the .eth extension are resistant to censorship because they are stored on decentralised smart contracts rather than traditional web servers. This makes them ideal for hosting content that could be removed on more conventional, centralised platforms. These sites can link to cryptocurrency wallets or IPFS-hosted files, further anonymising the group's digital presence.

Challenges in combating extremism in Web3:

i. **Censorship resistance:** Web3 technologies such as ENS and IPFS are designed to resist censorship, making it challenging for traditional regulatory measures to control the spread of extremist content.

ii. **Anonymity:** The anonymity platforms such as ENS, Matrix, and Monero provide complicates investigations into illegal activities and funding, making it difficult for authorities to trace the individuals behind the content.

iii. **Decentralised nature:** The absence of a central authority in Web3 platforms means there is no single point of control, making it hard for governments and law enforcement agencies to shut down or moderate harmful content.

How to investigate EthLinks

Since **.eth** is not a traditional domain name system (DNS) domain, it cannot be accessed directly using conventional domain resolution methods. However, **.eth** domains can be accessed using the **.link** extension, which acts as a bridge to retrieve the content stored on the ENS.

Here are some tips:

i. Use **Etherscan ENS Lookup** to check domain ownership, linked addresses, and past transactions.

ii. Example: Searching **'extremistdomain.eth'** can reveal the Ethereum wallet funding it.

iii. Use **Eth.Limo** to access ENS websites (e.g., extremistdomain.eth.limo).

iv. Use **IPFS gateways** to retrieve content if hosted on IPFS.

v. Use **Blockchair** to check Monero (XMR) or Bitcoin transactions from ENS-linked wallets.

vi. Monitor **Ethereum transactions** for suspicious fundraising or money laundering.

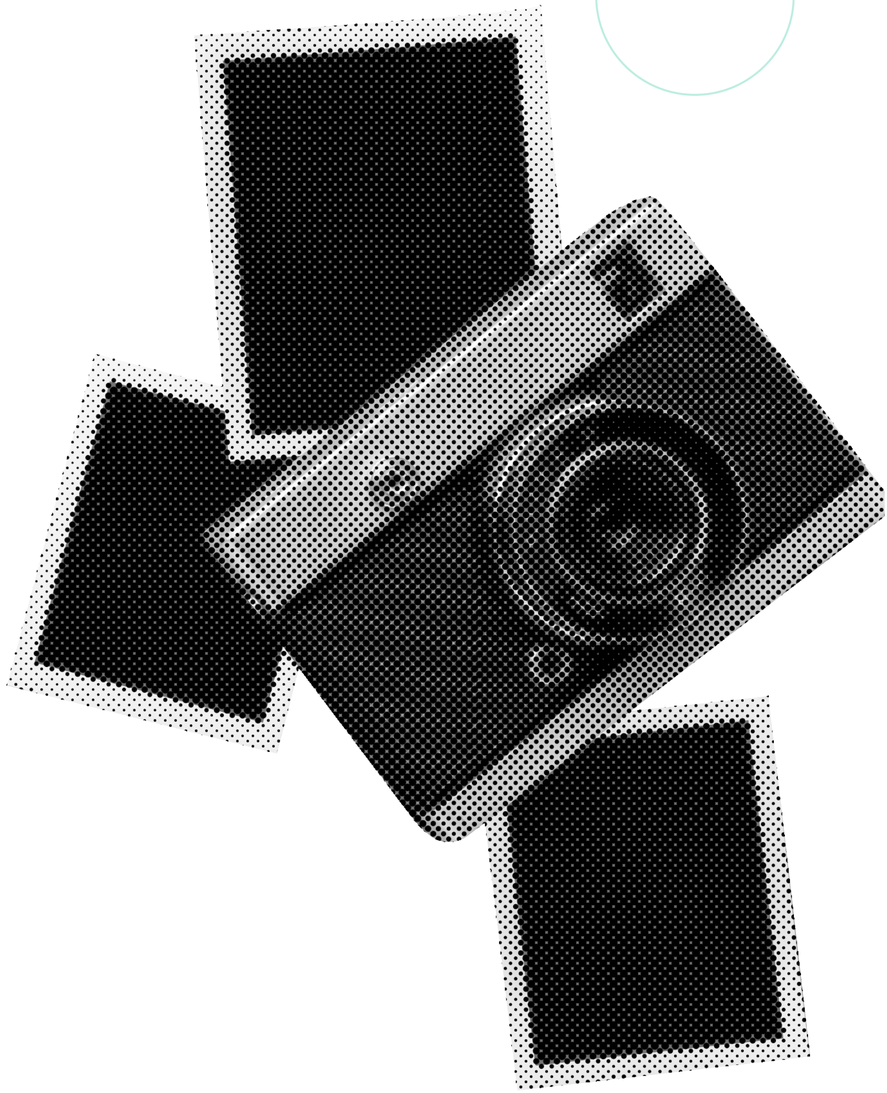
6.2 OSINT

6.2.1 Images

Images are often perceived as accurate representations of events or individuals, making them vulnerable to manipulation by bad actors who spread false narratives. They achieve this through context manipulation, digital alterations, or misleading captions. Journalists and CSOs need to possess the skills to critically analyse images and discern their true meaning and intent.

Common tactics bad actors use to deceive online audiences with images include:

a. **Alterations:** Digitally manipulating images to convey a specific agenda, message, or narrative.
b. **False context:** Sharing genuine images with incorrect contextual information, such as outdated photos or misleading locations.
c. **Missing context:** Deliberately omitting information when sharing an image.
d. **AI:** Using tools such as Dreamstudio and Midjourney to create AI-generated images.
e. **False attribution:** Placing fake quotes next to images of prominent individuals, often on digital cards.



Pillars of image verification

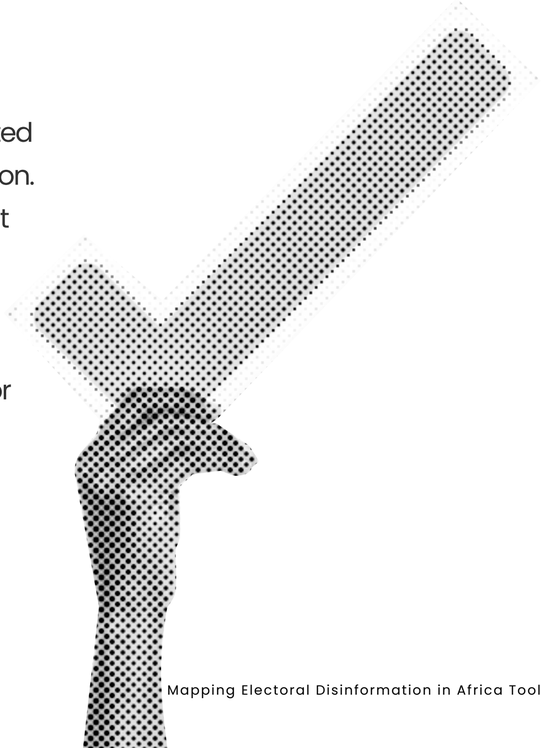
Determining whether an image is AI-generated, altered, or miscontextualised can be challenging. Therefore, it is essential to examine various aspects of the image to assess its authenticity.

The First Draft recommends five pillars for evaluating the legitimacy of an image:

a. **Provenance:** Are you looking at the original image?
b. **Source:** Who took the original image?
c. **Date:** When was the image captured?
d. **Location:** Where was the original image taken?
e. **Motivation:** Why was the image taken?

Other factors to consider when verifying images

a. **Check for red flags:** Look for inconsistencies such as mismatched fonts or misalignment between the image and text. AI-generated images often feature unnatural human traits, for example extra fingers or distorted facial features, which can indicate manipulation.
b. **Inconsistent edges:** Examine the edges of objects within the image. Excessively sharp, jagged, or uneven edges may suggest that elements have been poorly pasted or digitally altered onto the original photo.
c. **Mismatched lighting:** Evaluate the placement and direction of shadows and light in the image. Inconsistent lighting, such as mismatched light sources or shadow placement, may reveal that parts of the image have been manipulated.
d. **Poor quality:** A low-resolution or blurry image can often mask signs of manipulation, such as pixelation around altered areas. Poor quality can be a deliberate attempt to obscure changes or make the image appear more authentic.



Sensational topics: Images tied to sensational or emotionally charged topics, such as those designed to provoke anger, fear, or sadness, are more likely to be shared. This increased likelihood of sharing can be a red flag, as such content is often intended to manipulate emotions or spread mis-/disinformation.

Image verification tools

Relying solely on the human eye for image verification can lead to inaccuracies. It is essential to complement visual assessments with image verification tools. These include:

Google Lens/Google reverse image search

TinEye

Yandex

Labnol/reverse

DupliChecker

Fotoforensics

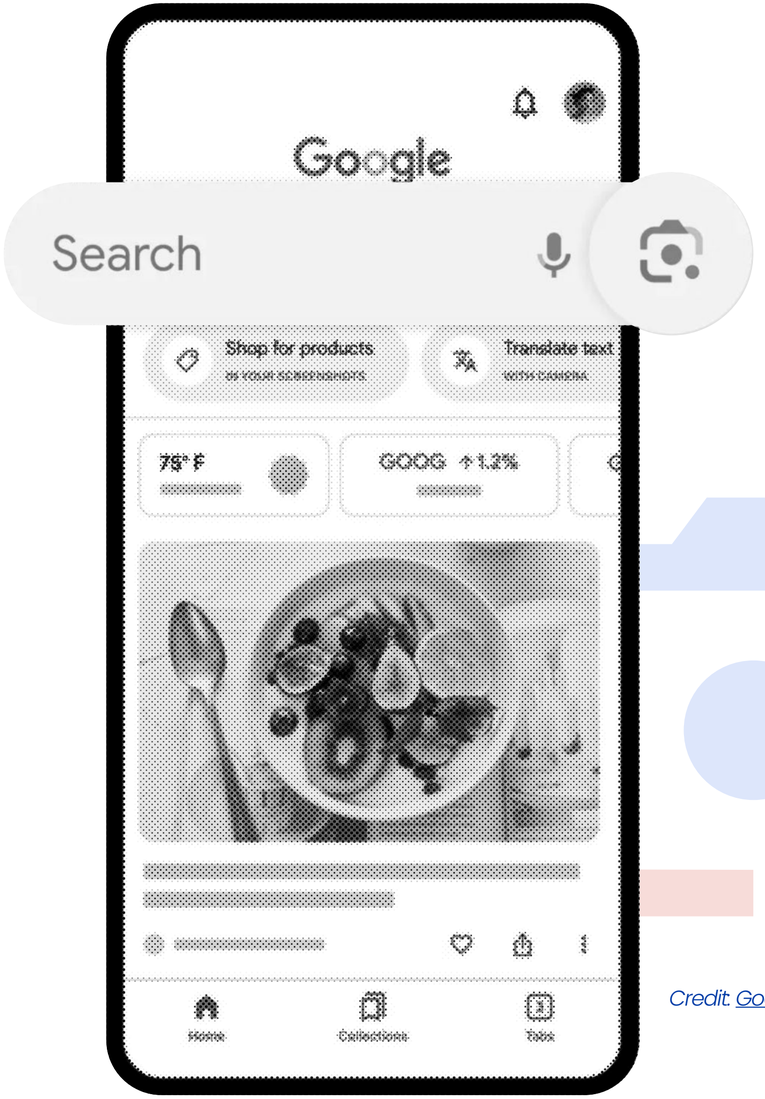
Hive Moderation

Forensically

Step-by-step guide to using Google Lens

Google Lens is an image recognition tool that identifies visual matches to uploaded items. Accessible via mobile, it helps determine the original use of images being fact-checked.

1. Download or screenshot the image you want to verify.
2. On the Chrome browser, tap the camera icon.
3. The icon directs you to Google Lens.
4. You can choose to either scan the image or upload it from your gallery.
5. Google will generate visual matches.



Credit: Google lens

6.2.2 Videos

Like photos, bad actors can easily manipulate videos, making them a prime medium for spreading mis-/disinformation. After all, seeing is believing – or is it?

How bad actors use videos to spread mis-/disinformation

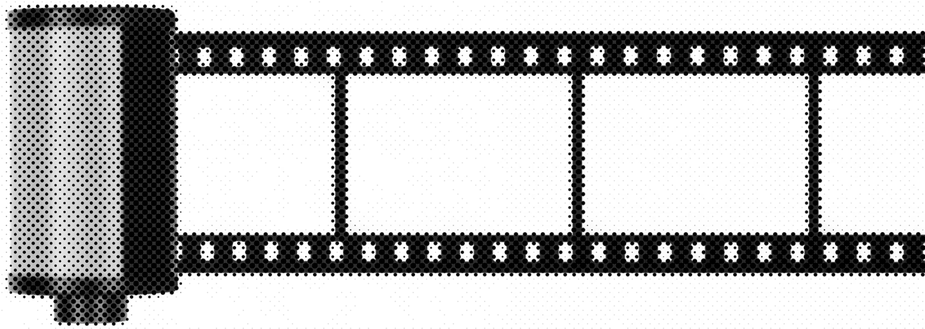
The tactics bad actors use to spread mis-/disinformation through videos are highlighted below:

a. **Video manipulation:** Altering a video to change its original meaning or perception.
b. **Missing context:** Passing off legitimate old footage as recent or using it in an unrelated context, often with misleading text.
c. **Deceptive editing:** Editing, cutting, or rearranging footage to distort its message.
d. **Malicious transformation:** Fabricated or doctored footage, such as deepfakes.

Importance of video verification

Video verification tools and techniques curb the spread of mis-/disinformation. They are important for the following reasons:

a. They help to differentiate fake from authentic information.
b. Ensure the accuracy of a video's content.
c. Help you avoid amplifying fabricated news and propaganda.
d. Add context, detail, history, and transparency to your stories.
e. Help find clues and corroborating evidence to verify videos.



How to verify videos

A video consists of a series of images played in sequence. To trace the origin of a video, follow these steps:

a. **Split the video into several still images or keyframes.**
b. **Run the separate stills through various search engines such as Google Lens, Microsoft Bing, TinEye, and Yandex.**

To implement the steps above, you should: pause the video you wish to investigate, capture several screenshots at different points, save them to your device, and perform reverse image searches using various search engines.

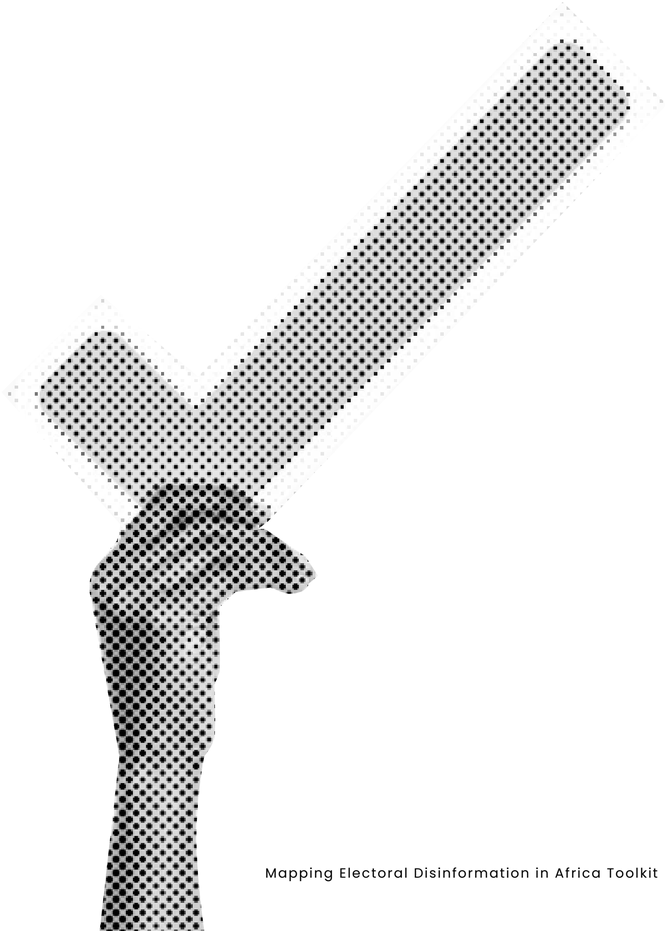
This process may seem tedious, which is where video verification tools become valuable.

One of the most commonly used tools for fact-checking video claims is InVID-WeVerify. The tool is designed to help verify the authenticity and context of online images and videos.

InVID-WeVerify's practical verification features

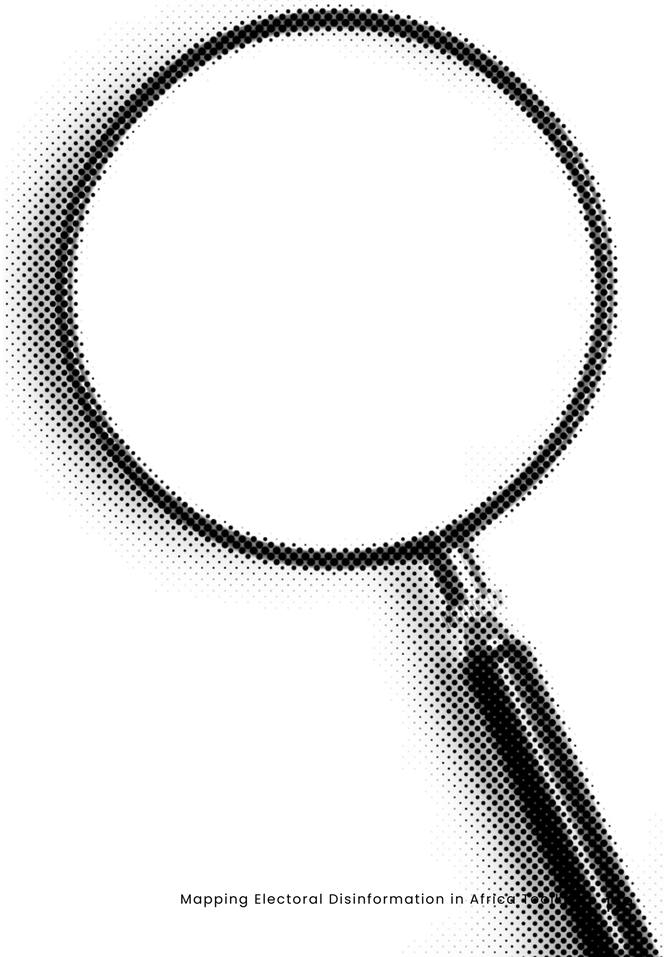
The InVID-WeVerify Chrome extension lets users fragment videos into keyframes and conduct multiple reverse image searches simultaneously, saving time and improving efficiency in fact-checking. The plugin offers a toolbox that allows users to:

a. **Quickly get contextual information on Facebook and YouTube videos.**
b. **Perform reverse image searches on search engines such as Google, Microsoft Bing, TinEye, and Yandex.**
c. **Fragment videos from various platforms, such as Facebook and X, into keyframes.**
d. **Enhance and explore keyframes and images through a magnifying lens.**
e. **Query X more efficiently through time intervals and other filters.**
f. **Read video and image metadata.**
g. **Check video copyrights.**
h. **Apply forensic filters to still images.**



Verifying videos using keyword searches

To debunk misleading videos, start by carefully listening to the content and extracting key phrases or statements. Use these to perform keyword searches across multiple browsers and video-sharing platforms such as YouTube. This process can help identify whether similar videos have been shared before in different contexts, which may reveal mis-/disinformation. By comparing the video in question with past instances, you can assess its authenticity, identify potential manipulations, or uncover misrepresentations, providing a clearer picture of its validity.



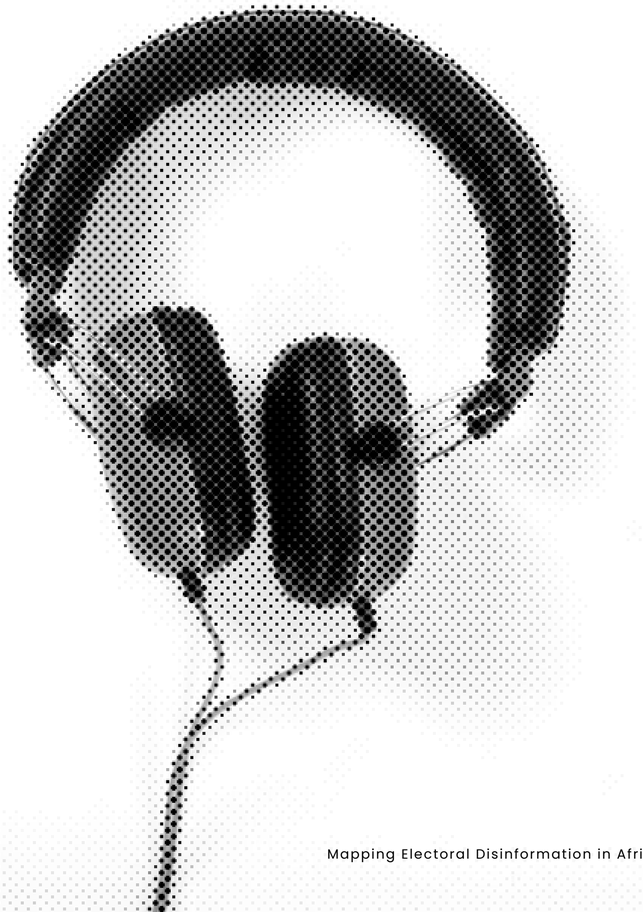
6.2.3 Audio

OSINT audio analysis is the systematic process of identifying, investigating, and verifying audio content to detect manipulation (e.g., deep fakes or splicing) or propaganda, using open-source intelligence tools and techniques.

It is important to investigate audio manipulation because fake audio can incite violence, sway elections, or damage reputations (e.g., deepfake CEO directives, forged political speeches). Tools such as Adobe Podcast AI, ElevenLabs, and Resemblyzer now make voice cloning more easily accessible to malicious actors, such as propaganda networks, scammers, and state-sponsored groups. As a result, addressing the problem also requires using AI-driven OSINT tools, especially because human ears often fail to distinguish synthetic audio from genuine recordings.

Focus areas include:

a. **Voice cloning:** Detecting AI-generated mimicry of real voices.
b. **Audio splicing:** Identifying edits that alter context (e.g., fake quotes).
c. **Synthetic audio:** Flagging entirely AI-generated voices.
d. **Amplification patterns:** Mapping bot-driven dissemination of suspicious audio.



Step-by-step guide to analysing audio

a. Collect audio samples.
b. Identify obvious signs of manipulation.
   i) Look for abrupt cuts, unnatural silences, or inconsistent background noise, indicative of splicing.
   ii) Use the spectrogram view of the audio recording and editing tool Audacity to detect AI-generated artefacts (e.g., overly smooth tonal transitions in Resemblyzer clones).
c. Compare suspicious audio to known voiceprints with **Resemblyzer AI.**

```
pip install resemblyzer
```

i. Install Resemblyzer.

```
from resemblyzer import VoiceEncoder, preprocess_wav
encoder = VoiceEncoder()

# Process reference (authentic) voice
ref_wav = preprocess_wav("reference_speech.wav")
ref_embedding = encoder.embed_utterance(ref_wav)

# Process suspicious audio
sus_wav = preprocess_wav("suspicious_audio.wav")
sus_embedding = encoder.embed_utterance(sus_wav)
```

ii. Generate voice embeddings.

```
similarity = np.dot(ref_embedding, sus_embedding)
print(f"Similarity Score: {similarity}")  # Scores range from 0 (no match) to 1 (identical).
```

iii. Calculate the similarity score.

6.2.4 Social network mapping

This method analyses social interactions to determine influence within a network. A social network map consists of nodes (individuals or groups) and edges (relationships like comments, reposts, or quote replies). The structure of these connections reveals influence levels, with larger networks indicating greater impact. On social media, such networks often facilitate CIB, which platforms prohibit. Understanding these patterns helps researchers identify and investigate manipulative narratives. Some open source tools that can be used to generate social network maps include:

Gephi: A network analysis and visualisation software package written in Java.
NodeXL: A free network analysis and visualisation software package for Microsoft Excel.

Step by step guide to network mapping:

a. Install Gephi from its website here

b. Once successfully installed, open the software and click ‘new project’ on the pop-up on screen or the ‘file’ tab.

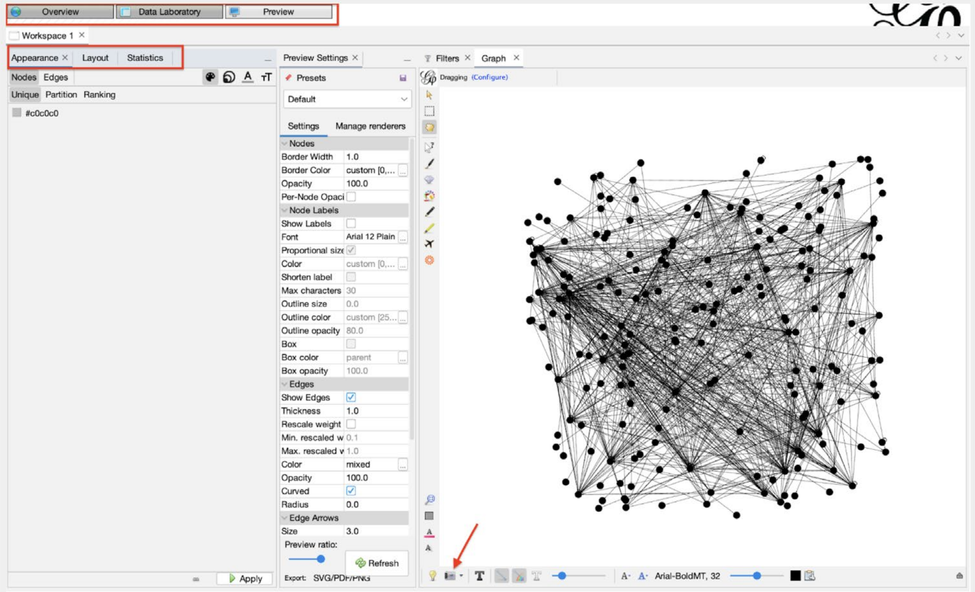
c. Gephi has three main interfaces; Overview (shows graphic analysis of the data that has been uploaded and allows for customisation of colours and graphics, the calculation of statistics, and the application of filters), preview (allows for additional visual customisation of the graph), and data laboratory (where the uploaded data is found. Here, you can further modify and inspect the data).

d. Import the preprocessed data into Gephi on the ‘file’ tab (File > Import spreadsheet).

e. Once you have confirmed from the data laboratory tab that the data is in your preferred state, shift to the ‘preview’ tab.

f. While on the ‘preview’ tab, calculate different metrics on the data using the ‘statistics’ option. Also choose a layout under ‘layout’. You can also modify the appearance of the nodes or edges – including the colour or size – to your liking, under the ‘appearance’ option.

g. Once satisfied with the graph, you can save the workspace by selecting ‘File > Save’. If you want to save a picture of the whole graph, click the camera icon at the bottom of the ‘Overview’ tab.



Screenshot showing a sample network map (Source: CfA via Gephi)

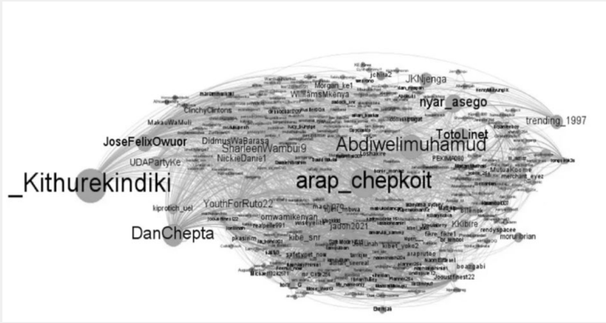
Real-life example: Artificially boosted hashtags during Kenya's 2022 elections

On 09 August 2022, when Kenya held its most recent election, there were two trending hashtags on X. The two main political factions used the hashtags #RutoThe5THPresident and #RailaTheEnigma concurrently with the hashtag #KenyaDecides2022 to sell their political agenda and campaign promises. The campaigns used bot accounts and keyboard warriors to artificially amplify these hashtags.

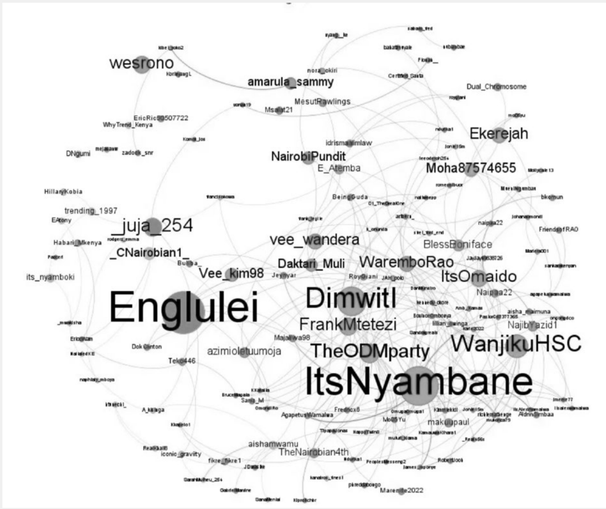
The hashtag #RutoThe5THPresident received ~16,500 mentions within 12 hours, with a primary network of 39 accounts contributing approximately 30% of all mentions, indicating a likelihood that different accounts amplified this hashtag through coordinated behaviour on social media.

Click here to read the ADDO article.

The #RailaTheEnigma hashtag received significantly fewer mentions, totalling ~3,510 on election day. Twenty influencer accounts and 30 bot accounts posted content using this hashtag. However, there was no strong indication of coordinated amplification as the top reposting accounts had reposted content 17 times.



A screenshot of a social network map showing clusters of X accounts using the hashtag #RutoThe5THPresident (Source: CfA using Gephi).



A social network graph showing clusters of X accounts using the hashtag #RailaTheEnigma during the 2022 Kenyan election (Source: CfA using Gephi)

Click here for more.

7.

Countering information disorder

7. Countering information disorder

This section covers Meta's and TikTok misinformation policies and their strategies for countering harmful content. It explores prebunking and debunking techniques, working with micro-influencers to combat hate speech during elections.

7.1 Community standards

Meta's misinformation policy

Meta classifies misinformation differently from other harmful content in its community standards. Unlike hate speech or violence, which have clear guidelines, misinformation evolves as new facts emerge.

Meta removes harmful misinformation that threatens public safety, election integrity, or political processes while limiting the reach of less harmful falsehoods like exaggeration or satire. It enforces transparency on AI-generated content, penalises coordinated deception, and restricts accounts that frequently spread misinformation, adapting its policies as digital threats evolve.

Other Meta interventions include:

i. Use of an informative label

Meta may assign altered media, such as computer-generated audio, images, or video that could mislead the public an informative tag. While some edits are harmless, Meta may mark deceptive alterations that have a high potential to misinform the public on critical issues or may decline them in ads.

ii. Meta community notes

In January 2025, Meta announced the launch of community notes, a user-driven fact-checking system inspired by X, replacing its third-party fact-checking. CEO Mark Zuckerberg said the rollout, starting in the US and expanding globally, aims to promote free expression and reduce moderation errors.

Community Notes contributors can submit explanations of incorrect posts by providing background information. Notes are published only if users with differing viewpoints agree they are helpful. Notes, authored and rated by contributors, must not exceed 500 characters, include a link, and follow Meta's community standards. Meta aims to clarify how diverse views contribute to the notes.Community notes will roll out in the US later this year. Early users can register on Facebook, Instagram, or Threads.

iii. Meta's third-party fact-checking (TPFC) programme

Meta's third-party fact-checking (TPFC) programme combats misinformation on Facebook, Instagram, and Threads, relying on independent fact-checkers certified by organisations such as the IFCN and the European Fact-Checking Standards Network (EFCSN), for rigorous, non-partisan verification. TPFC remains active globally, despite Meta's plans to roll out community notes.

Meta

Click here to view Meta's Community notes

TikTok misinformation policy

TikTok removes misinformation that threatens public safety, disrupts elections, spreads false health claims, denies climate change, or promotes conspiracy theories that incite violence or prejudice. This includes misleading crisis reports, deceptive electoral content, unproven medical advice, and harmful narratives targeting individuals or communities.

To counter misinformation, TikTok implements several measures in response to such content, regardless of intent. These include:

i. Content moderation by integrity and authenticity moderators, who consult a global database of fact-checked claims to assess content for the 'for you feed' (FYF).
ii. Partnerships with more than 20 IFCN-accredited fact-checking organisations evaluating content in more than 50 languages.
iii. Labelling content as 'AI-generated' to provide users with context.
iv. Redirecting searches to authoritative sources or adding informational banners.
v. Alerting accounts that violate misinformation policies or removing those that repeatedly share misleading content.

Click here to view TikTok misinformation policy

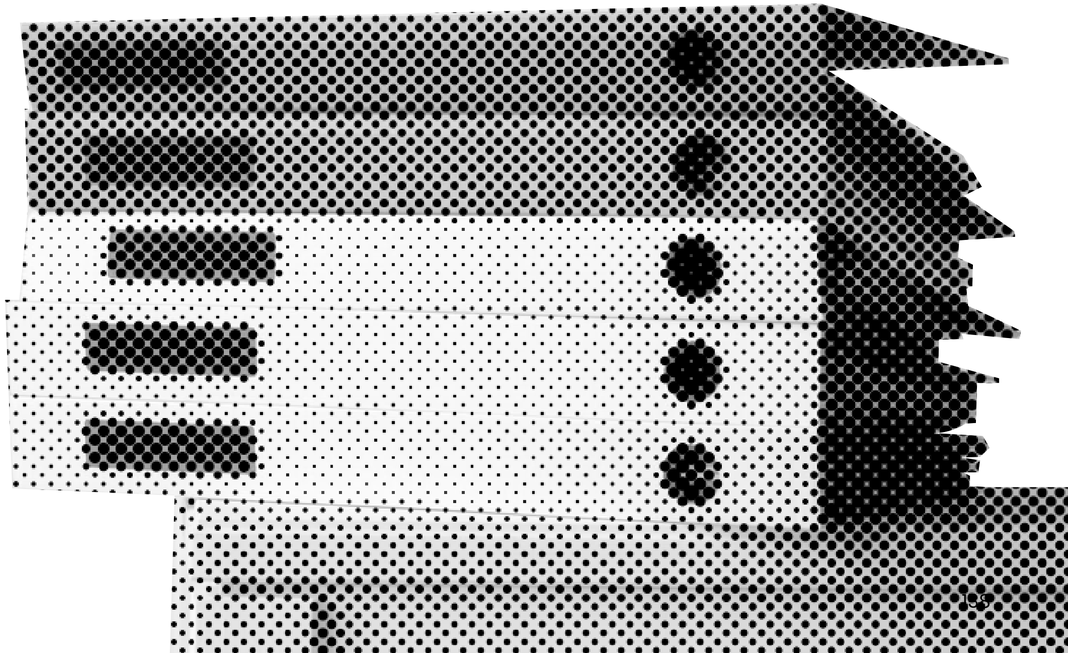
7.2 Debunking

Debunking involves verifying a claim's truth by assessing evidence, consulting credible sources, and providing a factual verdict. It explains why the information is false and addresses misconceptions fuelling misinformation. This process helps mitigate potential harm, especially during major events.

Debunking is crucial in countering misinformation by clarifying facts, providing context, preventing societal harm, and guiding audiences toward reliable sources.

Elements of a good debunk

a. **What is the claim?** Briefly describe the information being debunked in one sentence.
b. **Where was the claim published?** Identify the publication or source from which the claim originated.
c. **What does the source show?** Determine whether the source supports or refutes the claim.
d. **What sources were used to verify the claim?** List the references used to check the claim's accuracy.
e. **How is the claim classified?** Assess the claim on the truth meter – false, partly false, satire, missing context, false headline, or hoax.



7.3 Prebunking

In addition to debunking false information, prior interventions are necessary to build audience resilience against misinformation. One such intervention is prebunking.

Prebunking, or pre-emptive debunking, involves teaching people to recognise and resist manipulative messages before they occur.

Social psychologist William McGuire likens prebunking to a 'vaccine for brainwashing'. Drawing from medicine, it aims to prevent the spread of harmful information, helping people critically assess content rather than accepting it as fact.

Differences between debunking and prebunking are illustrated in the table on the right:

Debunking	Prebunking
Exposes a false claim.	Teaches people to spot false claims even before encountering them.
Primarily the role of journalists, fact-checkers, and professional mythbusters.	Focuses on empowering the audience to make better decisions in how they consume information.
Tends to focus on explicit misinformation, leaving out implied misinformation, which still has the possibility of misleading audiences.	Focuses on interventions that target the audience's behavioural patterns, thus curbing misinformation before it spreads.

Prebunking is vital in combating misinformation by exposing manipulation tactics and common false narratives, making audiences more resistant to deception. It also mitigates misinformation spread more effectively than debunking alone, as not all false claims can be flagged, debunking may push misinformation to other platforms, and flagging is often perceived as censorship.

What to consider when prebunking

a. **What information do people want to know:** Use tools such as Google Trends to identify trending questions or issues, as well as upcoming events such as elections or health campaigns. Then, ask: What confuses people? Which narratives might bad actors exploit? Help your audience recognise these tactics and narratives to avoid falling victim to misinformation.

b. **What techniques are commonly used to spread misinformation?** Misinformation spreaders use techniques such as character attacks, emotional manipulation, impersonating influential figures, polarisation, and promoting conspiratorial ideas to amplify their impact.

c. **Demonstrate using an appropriate example:** False narratives often resurface during events such as elections, disease outbreaks, or major legal rulings. Highlight the misinformation your audience is likely to encounter.

Important elements to remember for your prebunk

a. Lead with facts.

b. Keep it simple but not patronising.

c. Make it easy to share.

d. Plan for misinterpretation by offering resources where the audience can get more information.

Ultimately, the goal is to create/foster a healthier information environment!

Process of prebunking

a. **Choose your subject:** Identify the type of misinformation you want to address. Monitor various sites for potential misinformation about ongoing or upcoming events.

b. **Choose your audience:** Determine who would benefit most from this information. For example, prebunking initiatives related to elections can target the general public.

c. **Define the goal:** Decide whether you aim to teach a new skill, such as conducting a reverse search, or to shift attitudes and behaviours.

d. **Select the approach:** Decide if you want to focus on tackling a specific narrative or addressing the technique used, especially for recurring issues.

e. **Choose the format:** Select the best medium for delivering your message to your audience based on where you engage with them, such as blog posts, illustrations, status updates, or threads on social media platforms.

7.4 **Defusing** (mythbusters + peacekeeping)

For CfA, mythbusters are nano- or micro-influencers with strong grassroots engagement and trusted audiences. We carefully select them from diverse communities and interests to ensure their reach extends beyond traditional news consumers. They include everyone from locally relevant fashionistas and sports or music personalities to grassroots civil society or religious leaders. They can also include grassroots storytellers such as traditional knowledge workers, jalis, and imams.

Mythbusters play an important role in combating hate speech and misinformation by using their platforms within established guidelines for political messaging initiatives – to share specific peacebuilding messages and research-backed counter-narratives. They engage their audiences with peace-focused messaging, challenge divisive rhetoric, and promote responsible civic engagement. Through targeted social media campaigns, mythbusters help defuse electoral tensions and guide citizens toward credible information sources, fostering a more peaceful and informed election environment

How to work with mythbusters during elections

a. Adopt an audience-first approach

Leverage existing partnerships with major social media platforms to analyse and understand your target audience. This audience may include an entire country undergoing elections or specific administrative regions holding elections.

What to do to gain deeper insights, allowing you to design targeted and effective social media peace campaigns;

i. **Conduct behavioural analysis:** How does your audience engage with information?
ii. **Psychographic analysis:** What are their values, beliefs, and motivations?
iii. **Sentiment analysis:** What are their attitudes toward election topics?

b. Identify and vet influencers

Select mythbusters with strong engagement in your target communities. Choose them based on their credibility, alignment with the campaign's goals, and ability to effectively communicate with specific audience segments. Once you select the influencers, provide them with contracts, which may include an exclusivity clause so they do not partake in conflicting campaigns.

What to look for while recruiting the mythbusters:

i. No history of **incendiary language or hate speech**
ii. No **explicit political campaign affiliations**
iii. No **negative engagement patterns with their audience**

c. Develop key messages

Creating a clear content brief is an important step in a successful Mythbusters campaign. This document distills complex investigative findings or dossiers into a simplified, easy-to-understand format for fellows. It removes complex technical jargon—such as 'lexicon'—and replaces it with straightforward language that helps fellows create engaging content.

The content brief is shared with fellows regularly, ensuring they have updated, well-structured information to support their messaging throughout the campaign.

During the message development phase:

i. **Ensure the insights in the content brief aligns with narratives that promote peaceful discourse and improve the flow of factual information in the ecosystem during elections.**
ii. **Understand main themes that influencers need to share with their audience so that the dissemination plan aligns with the campaign's overarching goals and expectations.**

d. Provide guidance and support

Maintain open communication with mythbusters through dedicated relationship managers. Offer resources, real-time support, and training to help them navigate sensitive topics and ensure messaging remains consistent with the campaign's peacebuilding goals.

The relationship managers should:

i. Ensure the mythbusters understand the campaign messages,
ii. Guide mythbusters on sharing campaign messages in their personal style, as mythbusters know how to communicate with their networks better than anyone else.
iii. Monitor every social media post to ensure each message never strays off the mark.
iv. Directly with the implementing teams to share feedback, confirm strategy development, and refine campaigns

f. Assess the campaign impact

Consistent monitoring or tracking and reporting are crucial for evaluating the impact of mythbusters' peace campaigns. Regular reports will offer insights into audience reach, engagement levels, and the effectiveness of different platforms and content strategies. Analysing social media trends and interactions, will give the implementing teams insights into how messaging influences public sentiment, and assist to adjust strategies accordingly to maximise impact.

e. Develop key messages

Once the mythbusters have shared the messages on their social media platforms, they enforce them to drive awareness by resharing the content and providing commentary alongside the messages. Imagine seeing the same message expressed in unique ways by many of your peers in your network – it is authentic, relatable, and trustworthy, and this ultimately drives action.

Remember:

i. Mythbusters use their own language, text, tone, and voice in their messages. Do not dictate that.

ii. Messages need to remain educational, interesting, and pertinent throughout a campaign.

8.

Organising your project

8. Organising your project:

Learning concepts, tools, and frameworks is not enough without a strong team to apply them effectively. A solid organisational structure and operational workflow are crucial for accurate, timely responses to threats. A well-structured team ensures efficient information gathering, analysis, and dissemination, enabling clear communication and coordinated action. Clearly defined roles and streamlined processes allow experts to work seamlessly towards shared objectives.

8.1.1 What should your team look like?

Defending democracy requires a strong investigative workflow and the right team. A multidisciplinary team—comprising community engagement specialists, journalists, data analysts, and technical experts—must work systematically to uncover and expose threats. This structured workflow guides investigations, combining OSINT, commercial social listening tools, and traditional journalistic methods to safeguard democratic processes.

8.1.2 Workflow: Step-by-step

An effective investigative workflow follows a structured approach, integrating OSINT, commercial social listening tools, and traditional journalistic methods.

1. Radar (Initial monitoring)

Continuous monitoring identifies potential leads and filters relevant data. Analysts and automated systems track sources using OSINT tools (search engines, Shodan, Maltego, social media search tools) and commercial social listening platforms (Brandwatch, Meltwater, Talkwalker). General tools like RSS feed readers and Google Alerts help capture emerging trends.

2. Analysts (In-depth research)

Researchers and data analysts conduct detailed investigations, dentify patterns, and generate reports. OSINT techniques include data visualisation (Gephi, Palladio), data scraping (Beautiful Soup, Scrapy), and source verification tools. Commercial tools provide sentiment analysis, influence tracking, and bot detection. General tools like spreadsheets, databases, and note-taking apps support structured research.

3. Technical reviews (Verification)

Technical experts validate findings, assess reliability, and identify errors. Digital forensics tools (Autopsy, FTK Imager), network analysis (Wireshark), and metadata analysis (ExifTool) ensure accuracy. Commercial platforms help detect bots and analyse networks, while general security auditing and code analysis tools strengthen verification.

4. Insights (Interpretation)

Analysts, editors, and subject matter experts interpret findings and assess their impact. OSINT reporting tools (Google Docs, PowerPoint) and mapping tools (Google My Maps) help visualise insights. Commercial social listening tools aid data interpretation, while software like Tableau and Power BI enable advanced visualisation.

5. Editorial (Publication)

Editors and journalists refine reports for accuracy, clarity, and public dissemination. Fact-checking tools, source verification platforms, and real-time monitoring ensure credibility. General tools like content management systems, collaboration platforms, and writing software streamline the publication process.

6. Engagement (Post-publication)

Social media teams and community managers interact with audiences, monitor responses, and counter misinformation. Commercial social listening tools track sentiment, while social media management platforms and comment moderation tools facilitate engagement.

9.

Staying safe: Operational security

9.1 Threat/risk assessment

A comprehensive threat or risk assessment will help identify critical assets and potential dangers, mitigate risks, and ensure operational security. This process should be dynamic, regularly updated, and tailored to the specific context of your investigative work.

a. Identifying threat actors

Understand the adversaries who may seek to undermine, intimidate, or retaliate against investigators. These can include criminal networks, freelance digital mercenaries, ideological campaigners, state-affiliated stratcom units, and troll armies and bot networks.

b. Assessing threat vectors

Threats can manifest in multiple ways, including:

- **Digital threats:** Doxxing, hacking, malware, phishing, or spyware.
- **Psychological threats:** Harassment, intimidation, online abuse, or social engineering.
- **Legal and institutional threats:** Censorship, defamation lawsuits, or government pressure.
- **Physical threats:** in-person intimidation, surveillance, or threats of violence.
- **Reputational risks:** Identity theft, mis-/disinformation, or smear campaigns.

c. Evaluating risk levels

Each investigation or exposure carries a unique level of risk. Conduct a risk evaluation based on:

- **Likelihood of attack: How likely is it that adversaries will retaliate?**
- **Potential impact: What are the consequences of attack, exposure, or retaliation?**
- **Exposure level: Are investigators working anonymously or publicly?**
- **Mitigation capacity: What security measures are in place to counteract threats?**

A simple risk matrix (low, medium, high) can help categorise threats and prioritise mitigation efforts. A sample is shown on the right:

Threat type	Likelihood	Impact	Exposure level	Overall risk level
Threat 1	High	High	High	High
Threat 2	Medium	High	Medium	Medium

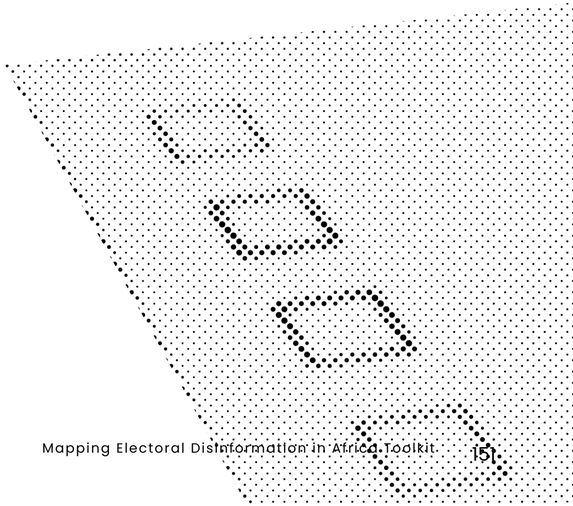
d. Implementing risk mitigation strategies

Once risks are identified, proactive steps should be taken to reduce vulnerabilities. These include:

- **Digital security:** Use encrypted communication platforms such as Signal or ProtonMail, virtual private networks (VPNs), and multi-factor authentication.
- **Social media hygiene:** Minimise exposure by limiting public personal information and using pseudonyms when needed.
- **Legal preparedness:** Know local laws regarding digital investigations and press freedom.
- **Operational protocols:** Work in teams and establish emergency contacts.
- **Mental health resilience:** Be aware of burnout, seek psychological support if needed, and establish peer support networks.

e. Continuous monitoring and adaptation

Threat landscapes evolve, so you must continuously update threat assessments, reassess security protocols and measures, and adjust strategies based on emerging risks.



9.2 Managing your research

Communication is a critical part of work as information travels from one person or place to another, but security of your information is not guaranteed by default. Due to the sensitive nature of some of the information you create, process, and store, steps must be taken to protect this information from getting into the wrong hands.

Here are communication protocols that ensure safety of information:

a. Use a VPN to protect your online activity from eavesdropping.
b. Meet new contacts in a public place when exchanging information.
c. Use burner phones to reduce the likelihood of eavesdropping.
d. Ensure you share documents with the right people by verifying their email addresses.
e. Regularly audit file sharing to ensure access is restricted to only those who need to see the file and revoke access once someone should no longer be able to see it.
f. Leverage secure communication tools to store and share information safely.

9.2.1 Secure communication

This list contains tools that are known to use encryption techniques and respect users' privacy. These ensure that no one snoops on your data or steals it when communicating.

Signal

Signal is a messaging app similar to WhatsApp, but with enhanced encryption. It provides a confidential means to exchange messages and make calls using end-to-end encryption. Signal also allows for the creation of groups where messages can be sent to multiple people, as well as group voice or video calls.

Jitsi

Jitsi allows users to have secure video conferencing and offers a free Jitsi Meets service on its website. You can opt to first create an account or simply use the free Jitsi Meets service. It uses encryption to secure communication.

ProtonMail

ProtonMail is a free and secure email service powered by a desire for improved privacy. It protects your communications from eavesdropping using end-to-end encryption and employs other techniques including two-factor authentication, anonymous account creation, phishing protection, self-destructing emails, and so on. Two good alternatives are Tuta Mail, formerly known as Tutanota, and Mozilla Thunderbird.

Virtual Private Network

A VPN protects your connection to the internet by creating a secure path for your internet traffic to flow through. This secure path ensures the privacy of your data as it moves across the internet. A VPN can also enable you to access websites blocked using geographical restrictions, as your traffic will seem to be coming from other locations. For example, a user in Kenya using a VPN can seem to be connecting from Germany, so if a website is blocked for Kenyan users but not Germans, this user with a VPN will be able to see the website. Two good examples are Outline VPN and Proton VPN.

9.2.2 Secure document and evidence

Documents and other files are sometimes used to store sensitive information. To protect this information from getting deleted or stolen, it must be stored securely and in multiple locations. The accounts used to share and store information should also be protected.

Here are some recommendations on how to save your documents and evidence securely.

a. Work in the cloud to ensure that you can access your documents through any device with an internet connection. Tools such as Google Docs and Sheets are free to use.
b. Protect cloud accounts with strong passwords and multi-factor authentication (MFA).
c. When using MFA, ensure you opt for an authenticator application, not SMS or email, which are dependent on network connectivity and are easy to hijack.
d. Adopt a three-way backup system, where your data is stored onsite, offsite, and in the cloud. Onsite refers to your office or on your work device, offsite means on a storage device that is kept away from your office, home (for remote workers) or in a secret location such as a safe. Cloud refers to online storage services such as Dropbox, Google drive, Keybase, and so on.
e. Be careful not to save documents on public devices that other people have access to. These include devices in a cyber cafe, library, or even ones shared with family members. Such devices are prone to damage, and file deletion is likely.

9.2.3 Secure storage tools

Keybase

Keybase provides a secure environment for exchanging messages, sharing files, and storing other sensitive information using encryption techniques. Keybase uses end-to-end encryption in its chat and its cloud-based storage called Keybase filesystem. Two good alternatives are Proton Drive and Onion Share.

Password manager

A password manager enables users to create strong passwords and store them securely using encryption. It can prompt users when a password has been exposed in a breach, among other features. This tool ensures you use strong passwords to protect the accounts you use in secure communication and file storage. The user is only required to remember one strong password called the master password, which grants access to the password vault. The master password holds all the keys to the kingdom so it must not be easy to guess or written anywhere. Two good examples are Bitwarden and Proton Pass.

CODE FOR AFRICA

MEDIA
MAPPING ELECTORAL DISINFORMATION IN AFRICA

Mapping Electoral Disinformation in Africa Toolkit