

Tik Tok and Manipulated Content: A Qualitative Analysis of Moderation Strategies to Misleading Political Campaign Messages in Kenya

Jemima Akoto¹, Catherine Muhatia¹, Mercyline Kerubo¹, Freda Agwa¹, Apolyne Anyango¹, *Prof Idah Gatwiri Muchunku²

¹Department of Journalism and Media, Multimedia University of Kenya

²Post Graduate Director, Multimedia University of Kenya

*Corresponding Author

DOI: <https://doi.org/10.51244/IJRSI.2026.1306000280>

Received: 13 June 2026; Accepted: 18 June 2026; Published: 04 July 2026

ABSTRACT

This paper investigates misleading political campaign material circulated on social media, especially in Kenya. It also investigates whether TikTok's content regulation strategies against politically manipulated information are effective. TikTok is of key interest because it is the most rapidly growing with increased fame especially among the Kenyan Youth. This study adopts a qualitative case study design informed by Gatekeeping and Social Responsibility for analyzing patterns. Twenty-one secondary data sources were utilized. In addition, Policy documents were scrutinized. More so, platform governance reports were considered. The manipulated content realized from the findings include fabricated videos, coordinated fake news campaigns, as well as algorithmically driven hate speech. The paper further outlines suggested moderation strategies, automated flagging of content, banning user accounts, as well as classifying deceptive content. The results reveal that there is still a lot that needs to be done. There are reasonable gaps in moderating Swahili-language and multilingual content. Successful intervention strategies require AI systems that are context-sensitive and collaboration with African civil organizations to develop structures that are locally relevant.

Keywords: Social responsibility, content moderation, gate keeping, TikTok, election campaign, political misinformation, Kenya

INTRODUCTION

1.1 Background of the Study

Social media has notably transformed political communication, offering political actors unprecedented access to mass audiences. Facebook, Twitter, Instagram, and TikTok have emerged to be key tools for shaping political stories, influencing voters, and informing public opinion (Chadwick, 2020, as cited in Nasebere, 2024; Kruikemeier, 2019, as cited in Nasebere, 2024; Gibson, 2020, as cited in Nasebere, 2024). Among these platforms, Tiktok is the most recent and fastest growing, with one billion users reached within five years of its launch (Capitol University Technology 2023). Tiktok's position in the social media space and its popularity among young users render it an obvious best target for examination. Hidalgo (2025) in Giuranno & Manni (2026) opines that Tiktok's rapid growth to the most used social media platform has spiked changes in the environment that concern modern elections. A recent study on the Ukraine political conflict indicated that on Tiktok, emotionally charged messages and videos that are misleading often outperform verified content.

Because of its short video content, it becomes an ideal platform for both misinformation and disinformation. TikTok users are prone to misinformation. Sensational headlines, language that exaggerated and emotionally charged content encourage impulsive communication, exposing users to unverified claims and distorted facts before they can confirm the authenticity of the information. Akmalia (2019) identifies several ways in which disinformation campaigns deepen social harm. They widen ideological divisions through the scattering of fake news speeches of hate, erode public trust in democratic institutions and the media, and intensify inter group

tensions increasing the risk of conflict. In response to the pressure from governments around the world, Tiktok has realized its contribution in spreading misinformation and has hence taken steps to mitigate its impact (Capitol University Technology 2023). This paper therefore sought to identify various forms of manipulated election campaign content, to find out the strategies used on Tiktok to control manipulated election campaign content, to evaluate the effectiveness of the Tiktok strategies to control misleading election campaign.

1.2 Statement of the problem

In Kenya, the communication of political campaigns has shifted to Tik Tok, where manipulated messages, hate speech, and coordinated disinformation campaigns have thrived with few regulations. According to Matiko (2025), TikTok is the most popular platform in Kenya where misinformation and disinformation are distributed. Although the platform has moderation policies, it is not well understood how they will work in the Kenyan context. This research paper thus aims to evaluate how well Tik Tok strategies in Kenya have been effective in abating misinformation in election campaigns.

1.3 Research objectives

1. To identify manipulated election campaign content on TikTok in Kenya
2. To identify control strategies used on TikTok to moderate misleading political campaign messages in Kenya
3. To analyze the control strategies effectiveness on TikTok to moderate misleading political campaign messages in Kenya

1.4 Research Questions

1. What forms of manipulated campaign content are identified by Tiktok in Kenya?
2. What control strategies are used by Tiktok to moderate misleading political campaign messages in Kenya?
3. How effective are the control strategies employed by Tiktok in moderating misleading political campaign messages in Kenya?

1.5 Significance of the study

This research adds knowledge and informs practice in a number of ways. To start with, it contributes to the knowledge of the operation of the moderation strategies of Tik Tok, and their failure in the Kenyan political communication case. Second, the results recommend practical solutions to policy makers, civil society, media regulators and organizations that deal with the managing online politically manipulated information. Third, the research reveals structural loopholes in the accountability of platforms that influence fact-checking organizations and digital literacy initiatives. Lastly, it can be used as a guideline to future researchers and it adds to the new area of research of digital political communication in sub-Saharan Africa. Findings from this study may also benefit policy makers, media regulators and other stakeholders by providing recommendations that can help in regulating and monitoring political misinformation on Tiktok. Further, the study will also reveal gaps and challenges faced by social media and other organizations for fact checking to help implement the right control strategies. This study also informs the field of knowledge in media studies, political and digital communication, also as a reference for future researchers.

1.6 Scope of the study

This study adopted a desktop qualitative content analysis approach to examine the effectiveness of control strategies used to manage campaign misinformation on TikTok in Kenya. The study focused on political campaign content shared on TikTok, including videos, comments, hashtags, and other forms of user-generated content related to political communication and elections in Kenya.

The research assessed the various strategies for moderation and checking facts, tactics used by Tik tok to regulate manipulated political content. In addition, the study examined secondary data documents such as

articles, policy documents, reports, as well as official statements that on misinformation intervention strategies on Tik tok in Kenya.

The research relied exclusively on documented sources without conducting any fieldwork study, surveys or interviews. Its core objective was to evaluate the effectiveness of current control strategies and identify the shortcomings in addressing manipulated political content on TikTok within the Kenyan context.

1.7 Limitations

This study was entirely based on desktop qualitative research. Data was drawn from online sources; the researchers did not directly interact with TikTok users or platform moderators. This limits the depth obtained from primary insights. Another limitation is that being an interpretive content analysis research, the researchers may have been influenced by their subjectivity. This study mainly focused on Tiktok and political campaign misinformation and therefore misinformation in other platforms was out of the study scope.

LITERATURE REVIEW

2.1 Introduction

The emergence of social media over a decade ago has brought about a paradigm change in the production, distribution, and consumption of political information (Castells, 2010; Jenkins, 2006). Billions of users of social media platforms like Facebook, WhatsApp, Twitter / X, Tik Tok, and YouTube are now able to create and distribute content to mass audiences in real time even without any formal training in journalism (Shirky, 2011). This democratization has clear benefits to the publics mostly in enhancing participation and information access. However, it has caused a new misinformation epidemic due to unprofessionalism in content creation, content distribution and gatekeeping (Wardle & Derakhshan, 2017) This has led to a contradiction of the existing media codes of ethics that guide the profession. The epidemic is more pronounced in Kenya due to the high penetration of mobile phones, extensive use of messaging apps like WhatsApp, and the lack of digital literacy (Ndlela, 2020; BBC Media Action, 2019).

This review of literature critically explores the role of social media platforms in the fight against misinformation with regard to unprofessional information creation, drawing from a body of peer-reviewed and quality policy research, including evidence from Kenya and beyond.

2.2 Theoretical Review

This study was grounded in the ideas conveyed in Gatekeeping Theory and Social Responsibility Theory, which are considered highly applicable to the research. This section briefly explains their key assumptions before examining their applicability, relevance in the study and limitations within the context of the study.

2.2.1 Gatekeeping Theory

Okeya and Asemah (2021) note that Gatekeeping theory formulated by Kurt Lewin in 1947 mainly argues that media corporations and specialists have a vital role in controlling flow of information from the rest of the community through emphasis and deemphasis and acceptance or rejection of information as noteworthy or unimportant. Based on the observation by Asemah, Nwammuo and Nkwam-Uwaoma (2017) in the words of (Bennett, 1996), gatekeeping theory gives the media the autonomy to keep some voices in while excluding others out of the public domain. Gatekeepers make these decisions based on inclinations, organizational practices, motives, common sense and newsworthiness. The theory proposes that media content goes through different channels before they make it to the reader or audience. The gatekeepers, ranging from institutions or establishments, editors, media owners and even advertisers, determine what content to publicize as well as how it is disseminated to the public (Asemah & Omula, 2013). Additionally, people also act as gatekeepers at the interpersonal level of communication when they select what to include or keep out of their emails, blogs or posts. Although the theory is applicable to various fields, including but not limited to politics and sociology, Lewin developed the thought while thinking about food choices before applying it to communication on mass media (Asemah, Nwammuo, & Nkwam-Uwaoma, 2017; Roberts, 2005),

Almakaty (2025) as well as Griffin, Ledbetter and Sparks (2022) contend that digital media has altered the traditional gatekeeping theory from human gatekeepers to users, platforms and algorithms serving as the keepers. Almakaty (2025) traces the evolution of gatekeeping from mass media specialists toward a coexistence of traditional and digital gatekeepers, increasingly shaped by algorithmic logic. Tech activist Eli Pariser's concept of "algorithmic gatekeepers" captures this shift: software filters now curate content based on web browsing history, geolocation, and user preferences, often invisibly determining what information reaches users (Griffin, Ledbetter & Sparks, 2022).

This study treated TikTok platform as an autonomous modern gatekeeper. While traditional gatekeeping entailed filtering and control through mass media specialists permitting, rejecting and foregrounding content (Okeya & Asemah, 2021), gatekeeping theory in this study helped explain Tiktok moderation strategies including but not limited to: reducing visibility; enacting penalties and suspensions on users; enabling user report systems and using AI to flag and remove inappropriate content. Gatekeeping theory was therefore instrumental in demonstrating how modern-day media keep voices in and out thus dealing with information overload and misinformation.

Owing to the nature of communication in the internet era, gatekeeping theory was somehow limited in meeting the research's third objective (assessing the success of TikTok's measures to control misleading election campaign messages. Roberts (2005) stated that gatekeeping theory in the digital media is slightly constrained as information may be filtered out of one platform only to resurface in a different site. Based on Roberts (2016), Cobbe (2021) observes that digital communication is extremely contextual, and decisions concerning content moderation are both complicated and ambiguous. This weakness impacted the study in its capacity to determine the effectiveness of the overall TikTok gatekeeping processes. Since the analysis was limited to TikTok, it could not be established whether the content that was moderated made its way to other platforms or reappeared in other formats in TikTok. This made it difficult to determine whether moderated content was prevented from reaching the public altogether. Research by Situma and Khejeri (2026) concluded that there is a prevalent use of memes, irony and sarcasm amongst Kenyan social media users, sometimes manipulated content may go undetected and fail to be flagged for moderation. This limited effective analysis of gatekeeping and content moderation regarding manipulated content.

Asemah, Nwammuo and Nkwam-Uwaoma (2017) assert that classical gatekeeping theory overlooked other socio-economic factors and gave so much power to editors or journalists who would control the process subjectively. Similarly, this study was unable to cater for manipulation that comes about owing to subjectivity arising from algorithmic bias and socioeconomic factors like platforms economic interests that subtly influence content moderation strategies (Cobbe, 2021). Cobbe (2021) observed that social platforms could be biased in their gatekeeping to protect their revenues which in the long run may affect objectivity. Consequently, this study incorporated social responsibility media theory to address gatekeeping theory's weaknesses regarding subjectivity in the control process.

2.2.2 Social Responsibility Media Theory

According to Obagwu and Idris (2019), the community suffers when valuable information is mismanaged, unreliable or withheld. While this study adopted gatekeeping to expound on the concept of selecting and controlling information, the study used social responsibility media theory to complement information control with service to the audience. This study was also informed by social responsibility media theory which was formulated by the Commission of the Freedom of Press in the mid-20th century to address the problems that libertarian theory brought forth. The theory's development was occasioned by a discovery by the Commission of the Freedom of Press that owing to press media freedom, the media hardly served the interests of the lower economic class, instead the affluent had more power thus creating the need for accountability and service to society. Additionally, the press was largely called out for bias and ethical issues like libel, slander and invasion of privacy (Asemah, Nwammuo & NkwamUwaoma, 2017).

Information flow within the media should be objective and enable intelligent and informed decision making on political matters and other life choices. The theory primarily advocates for a balance between independence and

duty to society, reminding media practitioners of their duty to avoid sensationalism and confusion (Asemah, Nwammuo & Nkwam-Uwaoma, 2017; Obagwu & Idris, 2019). Social responsibility media theory was useful in this study as it justifies the need for TikTok to effect content moderation strategies in a world of manipulated content and misinformation. The theory's emphasis that media should ensure accountability and service to society provides a strong justification for the study. Platforms are in a powerful position to ensure objectivity in communication even more than publishers, journalists or governments as platforms can: gauge news exposure and impact; use AI and algorithms to mitigate misinformation and trace the flow of misinformation back to the sources (Roberts, 2005; Rashidian et al., 2019; Vosoughi, Roy & Aral, 2018). Social responsibility media theory was instrumental in the study as it emphasized the need for platforms to take up and implement solutions to the research problem: prevalent use of deceptive and embellished messages on TikTok during election campaigns in Kenya.

Social Responsibility Media Theory complemented Gatekeeping Theory in two key ways. First, it explains why gatekeeping is necessary in a misinformation-saturated information environment. Second, it asserts that the gatekeeping process must be guided by public interest rather than commercial or political motives. From a study on the 2017 general elections in Kenya, Maweu (2019) observed that election periods in Kenya are characterized by prevalent use of social media to spread disinformation and cyber propaganda and to privilege certain voices and narratives over others. Additionally, as social media responsibility theory was developed to counter commercial media's influence and to counter the effects of media freedom which had eroded accountability and led to harm to the society by privileging certain narratives or voices, it was relevant in explaining the need to ensure objectivity and equality in communication in Kenya during election periods.

Although social responsibility provides ideal ethical frameworks to work with, in practice, digital environments inhibit effective implementation of the ideals. Traditionally, governments and media owners influenced what was considered public interest. Currently, platforms' "responsibility" is increasingly influenced by profits, the political environment, engagement metrics and even algorithms (Cobbe, 2021). It was therefore challenging to establish whether the actions taken for content moderation were independent and impartial.

2.3. Defining the Misinformation Problem: The Role of Unprofessional Content Creation

To understand the role of platforms in fighting misinformation, it is first necessary to define the phenomenon. In defining misinformation Aïmeur et al. (2023) offer a definitive global synthesis; defining misinformation as information communicated without intent to deceive, disinformation as intentionally false information and mal-information as true information intended to cause harm. Aïmeur et al. (2023) findings indicate that amateur user-generated content is the most common source of all three on modern social media.

Building on the findings of Aïmeur et al. (2023), Soliman (2024), in the *Journal of Information Technology*, engage with the concept of "two-sided disinformation platforms", in which the same structures that reward legitimate user engagement reward unprofessional creators for publishing and promoting disinformation. He contends that platform actions are often unethical in impact but shielded from direct liability by intermediary liability laws similar to US Section 230. He argues that this insulation eliminates one of the incentives for platforms to devote resources to combating unprofessional misinformation.

In countering unethical content, Gongane et al. (2022) provide an overview of the technical approaches used such as the processing of natural language, recognition of images, network analysis, and the review of human queues. They observe that as much as the moderation tools have become increasingly sophisticated, they are not well-tuned to the contextuality, multilingualism, and cultural specificity of references that are features of unprofessional content in non-Western digital media.

In their contribution to this discourse, Wang et al. (2025), show that algorithmic moderation systems are not able to account for creator intent, an important factor in determining the boundaries between satire, urban myths, and genuine misinformation in the amateur stream. They opine that intent-blind moderation leads to both excessive censorships of benign amateur content and insufficient censorship of harmful misinformation, with disproportionate impacts on creators from marginalized populations.

2.3.1 Platform Responses: Global Evidence on Moderation Strategies

According to Ali et al. (2023) in the Journal of Development Economics, social media platforms have adopted a multi-dimensional approach to misinformation at the global scale. They evidence this from a randomized control trial of the effectiveness of platform-adjacent information interventions through warning labels, accuracy reminders, and digital literacy prompts which are delivered at the point of content consumption. Their findings are that even short accuracy prompts largely decrease the sharing of untrue content hence suggesting the promise of behaviorally minded platform design as a supplement to punitive moderation. The authors however warn that the effects are smallest among users with the lowest initial digital literacy skills, the very group that is most likely to produce and share unprofessional content.

On Computers and Human Behavior Reports, Corsi et al. (2024) explore the growing trend away from platform-driven moderation towards community-led or crowdsourced moderation. Their Reddit analysis of spontaneous community moderation shows that peer-based enforcement of community norms can be highly effective at combating misinformation, while supplementing, rather than replacing the top-down platform moderation. Corsi et al. (2024) opine that for community moderation to be effective they require a sense of community, easy access to reporting tools, and responsive platform infrastructure that promptly addresses community reports.

The MIS (2026) Quarterly study on social media moderation and content generation offers empirical evidence on the effects of platform bans, which is one of the most severe moderation measures. Using large-scale data on banned users across platforms, the researchers find that bans temporarily reduce the creation of unprofessional content, but often displace unprofessional misinformation publishers, who move to less-moderated platforms. This finding underscores the shortcomings of punitive moderation practices in the absence of platform coordination and the need for more systemic solutions.

2.3.2 Misinformation on social media in Kenya: Scale, Platforms, and Key Actors

Kenya is a key case study for understanding misinformation in the rapidly digitizing Global South. Mudavadi et al. (2025) describe how this design characteristic enables the easy exchange of unprofessional voice notes, doctored images, and republished misinformation in Kenyan WhatsApp groups. Mudavadi et al. (2025) found that over 83% of Kenyan WhatsApp users surveyed had encountered misinformation on the platform. The low cost, group-based structure and a high level of social trust of WhatsApp makes it the most dominant channel of sharing news at community level in Kenya. However, the very same aspects, in this case, its end-to-end encryption, makes automated moderation much harder than on publicly accessible platforms.

McKay (2022), in an analysis for the Stimson Center, offers a detailed description of the 2022 electoral disinformation ecosystem in Kenya. This includes the emergence of a semi-professional "disinformation industry" of paid bloggers, coordinated influencer networks, and citizen volunteers mobilized to produce and share disinformation. McKay notes the low economic threshold to join this industry, demonstrating how the ease of use of content creation tools on platforms such as Facebook, Twitter/X, and TikTok has enabled unprofessional political misinformation to become a profitable industry.

The CIPESA (2022) report on disinformation in the Kenyan political space provides a detailed mapping of this ecosystem, documenting the actors (politicians, political party officials, paid influencers, and the general public), channels (mainly WhatsApp, Facebook, and Twitter/X), and consequences (hate speech, voter suppression, ethnic tensions) of concerted disinformation campaigns during the 2022 general election period. This report is critical of the platforms' enforcement actions, citing cases where the platforms took long to remove content that was in violation of community standards. This was due to slow human moderation and insufficient automated content moderation for Swahili language and code-switched content.

Oloo Ong'ong'a (2021) using qualitative interviews on Kenyan youth, provide an important age and class dimension, which shows that the Kenyan youth are overrepresented as both creators and consumers of fake news on new media platforms. The findings indicate that platforms like TikTok, Instagram, and YouTube's are tools for rewarding creators by monetization of content, follower counts and/or algorithmic recommendations. These platforms act as structural incentives for many young unprofessional creators to produce sensationalist, often misleading content for the sake of social capital and revenue.

2.3.3 Media Professionals, Fact-Checking, and the Limits of Platform Partnerships

A key theme of research in Kenya concerning unprofessional media content in platform dominated information ecosystems, examines the challenges faced by professional journalists, fact-checkers, and media literacy educators. Matanji et al. (2024), using qualitative interviews with journalists from Kenya and Senegal, recommend that fact-checking and media literacy operate as parallel but should complement each other; the former as a reaction to existing misinformation and the latter as a prevention of misinformation. An audience study by Madrid-Morales et al. (2025) identifies a conceptual gap between how media professionals and Kenyan users understand and respond to misinformation. Whereas professionals conceptualize misinformation in terms of fact-checking and source verification, many people who use WhatsApp as their primary news source tend to rely on personal trust and community reputation, rather than professional journalistic practices. This gap has far reached implications for the effectiveness of platform-based fact-checking labels and warnings, which assume a degree of media literacy and trust in institutional fact-checkers that is not yet common among Kenyan users.

In Kenya, the "Social Truth Queries" research (2024), offers an alternative platform-integrated approach to combat misinformation by leveraging social network structures to crowd source truth assessments. Initial findings indicate that truth signals from the local community can be more effective than algorithmic filtering and professional fact-checking in certain circumstances. The study thus suggests a blended approach to platform governance that harnesses social truth while preserving platform authority. This system is especially applicable to contexts with a high degree of unprofessional content in languages and cultures that are difficult for algorithms to process.

2.4 Structural Critiques, Accountability Gaps, and Recommendations

The accumulative weight of the evidence examined leads to a structural critique of social media platforms' contributions to misinformation: their business models are misaligned with the effective mitigation of unprofessional content-driven misinformation. Aïmeur et al. (2023) describe this as the attention economy: platforms make advertising revenue from user attention, which is attracted by emotionally stimulating, often inaccurate content. Soliman (2024), suggests that legal impunity compounds this misalignment, enabling platforms to externalize the social costs of unprofessional misinformation onto communities, democracy, and public health, while capturing the advertising revenue generated by high engagement with misinformation.

CIPESA (2022) and McKay (2022) developed this argument to the political economy of platform investment in Africa. They indicated that despite the advertising revenue they draw from Kenyan and other African markets, platforms have shortcomings which include disproportionately small local moderation teams, little investment in African language AI training data, and only engage superficially with local civil society and regulatory agencies.

Under investment is of importance in the context of Kenyan online spaces, where the prevalence of multi-lingual code-switched content (Swahili, English, and vernacular languages mixed) pose a challenge for automated systems trained primarily on English data.

A recommendation by Gongane et al. (2022) and Wang et al. (2025) is that platforms should build and roll out more context-aware, intent-sensitive moderation systems, including specific investment in data to train AI systems in Global South linguistic and cultural contexts. Ali et al. (2023) and Corsi et al. (2024) collectively recommend a hybrid moderation approach that integrates platform-level AI tools, community-level norm enforcement, and behaviorally informed user-level moderation interventions, including accuracy prompts and labels. The MIS Quarterly (2026) paper calls for inter-platform coordination to stem the flow of banned unprofessional creators to permissive environments, concluding that platform-based moderation will never be enough so long as the broader information environment offers permissive options.

In terms of policy recommendations for Kenya in particular, Matanji et al. (2024) and Madrid Morales et al. (2025) propose that platforms should invest more on financial and technical support for locally-based organizations such as Africa Check and Code for Africa, which have the expertise, language skills, and trust that global platforms do not. In particular, Mudavadi et al. (2025) recommend that WhatsApp and Meta create Kenyan versions of message forwarding limits, local literacy initiatives, and community ambassador programs

as network level structural solutions to unprofessional content. The Social Truth Queries study (2024) offers community-validated fact-checking as a scalable, localized complement to such platform-based moderation.

2.5 Conclusion

The studies that were analyzed in this chapter present a paradoxical situation that is quite disturbing and consistent at the same time. Much as social media platforms contribute to the dissemination of misinformation, it is their task to mitigate it (Gillespie, 2018; Suzor, 2019). In Kenya, this strain is worsened by strong mobile penetration, WhatsApp control, politically driven disinformation networks, multilingual content space, and digital illiteracy.

The digital platforms have incorporated a number of moderation strategies like bans and down-ranking, but the indications are that these responses are still reactive, culturally blind and technically wanting. More importantly, they are structurally weakened by profit driven engagement models that support the same content that they purport to bury.

For this crisis to be solved, it requires context-aware AI moderation, actual investment in African civil society and fact-checking groups, community-based governance structures, cross-platform coordination, and strong media literacy initiatives inherent in Kenyan education systems.

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the methodology used to conduct this the study. It describes the research design, the site of study, the targeted population. It also presents the sample size, sampling frame, and sampling techniques. It further describes the methods, instruments and procedure for data collection. It states some ethical considerations and pilot study procedures. It explains the strategies use to analyze and present data.

3.2 Research Design

A qualitative research design was adopted to allow for an in-depth examination of manipulated political messages on Tik tok, in the Kenyan context. A qualitative approach was appropriate because this study sought to understand how TikTok moderates politically manipulated content in Kenya, a question that requires interpretation of meaning, context, and institutional practices rather than numerical measurement (Henn et al., 2006; Creswell, 2014). Phone call interviews was also done to gather information based on research objectives. The study was based on the interpretivist paradigm, which rests on the argument that reality is socially constructed. This social construction is shaped by human interaction, communication patterns, and institutional frameworks. From this angle, misleading political content on TikTok were understood as socially constructed and shaped by political players, platform governance structures, algorithmic systems, as well as user engagement. This approach was a fit for context-based nature of social media political communication on Tik tok (Creswell, 2014).

The study specifically applied a qualitative case study design that focused on TikTok as a famous online platform within Kenya's political communication landscape. The use of a case study method made it possible to examine of current issues where conditions where real-world issues are intertwined with the phenomenon under investigation (Yin, 2018). Similarly, Creswell (2014) opines that qualitative case study is an in-depth study that utilizes multiple sources of data to generate trustworthy results.

3.3 Context of the Study and Population Under Investigation

3.3.1 Context of the Study

This study was conducted within Kenya's digital communication space, with TikTok serving as the primary online setting for study. The Tik tok platform functioned as a space where political communication, user engagement and moderation activities took place. The study focused on politically misleading content

circulating on the platform that is associated with Kenyan and related hashtags used in electoral and key political moments.

3.2.1 Target Population

The target population included online content and moderation strategies related to political misinformation on TikTok within the Kenyan space. The study was mainly centered on political campaign content, including captions, videos, hashtags and other audiovisual materials. The study also explored TikTok's moderation and governance systems which include community regulations, policies on misinformation and election integrity frameworks as well as and transparency documentation. Target population was also key informants from expert bodies for in-depth phone interviews.

3.4 Sample Selection Frame

Samples were drawn from freely accessible TikTok's politically manipulated content. It included online materials misleading political messages such as captions, videos, hashtags, manipulated visuals and as well as distorted audio clips. To achieve the first objective, the sampling frame consisted of posts with false political claims, deceptive captions, altered visuals, outrageous political footage, fake identity and false narratives. Concerning the second research objective, misinformation interventions, community guidelines, political integrity and transparency frameworks were included as Tik tok moderation documents in the sampling frame. The sampling frame for the third objective encompassed political campaign content that had been subjected to been flagging, removal, labelling and restriction. These materials revealed whether TikTok moderation strategic interventions had been effective. For the key informants, 5 to 10 key informants are usually sufficient to make informed generalization.

3.5 Sample Selection and Sampling Procedure

Participants were deliberately selected according to predetermined criteria that was relevant to the objectives of the study. Purposive sampling was used to establish content-rich data sources directly connected to the research questions. This strategy was relevant because it enabled the purposeful selection of materials that gave comprehensive insights (Creswell, 2014). Criterion sampling was used to select TikTok political campaign content that met specific characteristics such as manipulated visuals, misleading political claims, deceptive captions or evidence of moderation interventions. Key informants were purposively sampled, base on expertise to deliver required information.

3.5.1 Size of study sample

Rather than being fixed in advance, the final sample was informed by data saturation and concluded when further data ceased to contribute meaningful new findings (Guest et al., 2020). The study analyzed approximately 40 to 60 TikTok political campaign posts identified as manipulated or subjected to moderation interventions. Relevant TikTok governance documents were also included in the interpretation of findings. 7 key informants were selected for the telephone interview. journalist, one person from pesa check, one political analyst, two researchers and two Tik tok users.

3.6 Tools of Data Collection

Data generation was carried out using selected research instruments including a qualitative content analysis guide and a document analysis checklist. The content analysis guide was used to identify and categorize manipulated political content and moderation interventions. The document analysis checklist facilitated systematic review of TikTok's policies and governance frameworks.

An interview guide was developed based on the study objectives. Telephone interviews were thus conducted with key informants drawn from media practitioners, fact-checking organizations, digital communication experts, researchers, and professionals knowledgeable about political communication and misinformation on social media platforms.

3.7 Data Collection Procedures

Data collection was conducted on TikTok political content which was identified using election-related keywords, hashtags and trends. Selected posts were documented through screenshots, transcription and field notes. Phone call interview with key informants was also done.

3.8 Ethical Framework

The conduct of the research was guided by established ethical requirements to ensure integrity and responsible practice of the investigation process. Although publicly accessible TikTok content was used, identifying details were anonymized to minimize potential harm and avoid amplifying misinformation.

3.9 Pilot Test

To refine the research process before the main phase of data gathering, pilot testing was undertaken to evaluate the clarity and reliability of the research instruments. The process involved a limited sample of Tik tok content alongside two to three key informants.

3.9.1 Reliability of Data Collection Tools

The ability to produce credible, accurate and findings that address the research objectives determined the reliability of instruments (Creswell, 2014). It was reinforced through detailed document assessment and peer review to ensure consistency.

3.9.2 Validity of the Research Instruments

Validity was concerned with ensuring credible and trustworthy findings (Creswell & Miller, 2000). This was enhanced through the triangulation of triangulation method of establishing frequencies as well as, prolonged engagement with the data, integrating from different sources as well as peer debriefing. These reinforced the overall interpretation and understanding of the research problem.

3.10 Data interpretation and Presentation

Thematic analysis was majorly used to interpret the identified information from interviews, based on Braun and Clarke (2006) ideas where patterns and themes emerged through repeated and consistent engagement with the data. Themes were arranged in codes. Meaning was developed through building of themes.

Qualitative content analysis was applied to classify the forms of misleading political content and moderation approaches. Policy documents were equally analyzed thematically. Findings were presented in a narrative form, supported by policy document references.

3.11 Chapter Summary

The chapter outlined the processes followed during the research. It covered aspects such as the context of the study, research design that included how selection of participants was done, as well as how data was collected. Ethical requirements were considered. Testing instruments and measures used to strengthen reliability and validity were also outlined. In closing, the chapter explained the analytical and presentation procedures that guided interpretation of the findings, which are discussed in the subsequent chapter alongside existing scholarship.

FINDINGS

4.1 Introduction

This chapter provides the summary of findings from the in-depth interviews with key informants. The chapter also provides the results of a thematic analysis of 21 secondary sources about the role of TikTok in the political communication landscape in Kenya. The results are structured based on the three research questions:

determining the types of manipulated election campaign content on TikTok; analyzing the moderation policies of Tik Tok; and assessing their efficacy in the Kenyan setting. It is noted that social media platforms give a forum for people to speak and have proven to be an effective tool for persuasion, particularly in political matters (Jenkins, 2006; Castells, 2010; Chadwick, 2013). Because of its high level of interactivity, it provides a fertile ground for manipulation and false messaging with an aim to sway people's political stands, perceptions, and behavior (Njagi, 2025). TikTok's dynamic design encourages active engagement, as online media platforms have a big impact on the information environment. According to reports, Kenya's social media penetration rate in the first quarter of 2024 was 23.5% (Njagi, 2025).

Njagi (2025) discovered that the dissemination of disinformation in Kenya occurs predominantly online, with organizations, people, and influencers creating fictitious social media accounts for ulterior objectives. They do this by frequently utilizing many accounts with different names, making it difficult to trace the originator.

4.2 What are the Various Forms of Manipulated Election Campaign Content on TikTok?

A study conducted by Madung (2021) under the Mozilla Foundation found that fraudulent political content was reaching millions of Kenyans via TikTok, and that videos of fake newspapers, TV bulletins, and polls were permitted to spread without any controlled regulation (Madung & Obilo, 2021). In a sample of 130 TikTok videos, Madung discovered that videos from various accounts containing political disinformation regarding the impending election, including distorted content, hate speech, and appeals to violence, had received 4 million views. Furthermore, the study revealed that disinformation mostly comprised slickly produced video content and attack ads making misleading claims about candidates, as well as threats to various ethnic populations. According to Madung, the information received amplification from the TikTok for You page algorithm, as many of the videos targeted an audience that was significantly larger than their followers. This is because when consumers start to appreciate a video, the TikTok algorithm prioritizes it in the "For You Page, exposing it to a larger audience (Zhou,2024).

A video falsely shared by William Ruto and captioned I hate the Kikuyu people who live there and want vengeance, according to Madung and Obilo (2021), got close to half a million likes - an example of how hate speech can go viral with almost no intervention by the moderator. In a different scenario, a video of a detergent ad was altered to have the narrator unfavorably describe and talk about the 'removal' of tribes like Kikuyus, Luhya, Luos, and Kamba. This video was accompanied by gruesome material from prior election violence to stir political tensions.

In a different study done by Amnesty International during the 2024 Gen Z protests in Kenya, it was discovered that different types of information were being distorted to malign the Gen Z movement. The government-sponsored disinformation campaign methodically sought to defame activists and undermine the movement's legitimacy. Amnesty International (2025) found coordinated operations by state-sponsored trolls and paid bloggers, known as '527 bloggers,' employed to distribute false material and smear Gen-Z activists. These influencers actively promoted pro-government messages and generated counter-hashtags to drown out protest messages (Amnesty International, 2025). For example, when demonstrators popularized *#RutoMustGo*, pro-government influencers rapidly responded with *"RutoMustGoOn"* to manipulate trending topics and minimize protest visibility.

The emergence of information disorder has made it difficult to discriminate between fact and fiction. 75% struggle to distinguish true from fraudulent news online (Mboya, 2024). This disparity between high social media news consumption and low media literacy opens up vulnerability that political disinformation exploits. It also underlines why platform-level moderation as a standalone is insufficient. This vulnerability led to manipulation since many people lacked the necessary digital literacy required to critically scrutinize information sources before believing. The profit-making misinformation business in Kenya worsened the situation, as many influencers used their competence to benefit from distributing political misinformation (Madung & Obilo, 2021; Mboya, 2024).

The phone call interviews revealed that manipulated political campaign content on TikTok appears in diverse forms. For example, participants consistently identified fake news, doctored videos, manipulated images,

misleading captions, impersonation of accounts as well as coordinated disinformation campaigns as the most common forms of misinformation circulating on the platform.

Most respondents observed that edited video clips are particularly effective in misleading audiences because they create a false impression of political events while appearing authentic. Others noted that political actors frequently use emotionally charged narratives, ethnic appeals, and misleading political claims to influence public perceptions during election periods.

4.3 Role of Social Media in Combating Disinformation

Efforts to combat disinformation on social media operate at three levels: user-level interventions, platform-level interventions, and government-level interventions (Gupta et al., 2022). Each level carries distinct strengths and limitations, particularly in the Kenyan context. Digital media literacy is commonly known as a preventative instrument at the user level. According to Matanji et al. (2024), in Kenya, fact-checking agencies and media literacy teachers have a hard time refuting fake news since, by the time a correction is posted, fake news has already been shared thousands of times in closed groups that cannot be accessed by platform moderators. At the platform level, Ali et al. (2023) discovered that accuracy cues and warning labels can decrease the sharing of fake content, but with the lowest results among users with low digital literacy - the very group most vulnerable to being exposed to manipulated TikTok content in Kenya. On the governmental level, Gupta et al. (2022) state that laws related to AI-generated content and deepfakes are being developed worldwide, yet they are not enforced yet. The regulatory aspect of platform accountability in Kenya is underdeveloped in particular, and moderation is left to Tik Tok itself.

Following the phone call interviews, participants identified several moderation strategies employed by TikTok to address misinformation and manipulated political content. These include artificial intelligence-based content detection systems, user-reporting mechanisms, content removal procedures, account suspensions, warning labels, visibility reduction measures, and partnerships with independent fact-checking organizations.

Most respondents on call acknowledged that TikTok has established formal community guidelines and election integrity policies aimed at addressing political misinformation. However, they observed that implementation remains inconsistent and inadequate within multilingual contexts such as Kenya.

4.4 What is the Effectiveness of the TikTok Strategies to Control Misleading Election Campaigns?

Tik Tok has established a number of moderation mechanisms to deal with the misleading political content, such as AI-based content flagging, removal of Community Guidelines breaches, labeling election integrity, and banning users (Capitol Technology University, 2023). The platform has also a Creator Marketplace policy which limits political advertising and labels with fact-checks in collaboration with third-party organizations during election times.

Regardless of these, the Kenyan experience indicates that the moderation of Tik Tok is mostly weak in Kenya. Madung and Obilo (2021) discovered that the videos that included disinformation, hate speech, and ethnic incitement garnered millions of views before being taken down. CIPESA (2022) explained this failure in part by the slow human moderation of Tik Tok and the lack of automated Swahili-language and code-switched content moderation systems. According to

Gongane et al. (2022), moderation tools trained on Western data, which are mostly based on the English language, always fail in multilingual African online environments. This is a structural weakness that Tik Tok is yet to sufficiently respond to in its investment in African moderation capacity.

The results thus suggest that the moderation policies that Tik Tok currently has in place are reactive, culturally blind to the Kenyan multilingual political communication context, and structurally flawed through the profit-driven engagement model of the platform (Soliman, 2024).

Even punitive measures like bans, as the MIS Quarterly (2026) study discovered, only tend to push out disinformation creators instead of eradicating them. The context-sensitive AI systems, long-term investment in

the local-language moderation, and active partnership with Kenyan civil society institutions, including Africa checks and Code for Africa, will be necessary to successfully mitigate political misinformation on Tik Tok in Kenya (Matanji et al., 2024).

A concerns the effectiveness of control strategies, interview respondents generally considered TikTok's moderation strategies to be only moderately effective. While participants acknowledged that harmful content is sometimes removed or restricted, many argued that moderation interventions often occur after content has already reached large audiences. Further, several respondents described TikTok's moderation approach as largely reactive rather than preventive. They observed that misleading political content frequently accumulates significant engagement before being flagged or removed.

CONCLUSION

This study examined manipulated political campaign content on TikTok in Kenya and evaluated the effectiveness of the platform's moderation strategies in addressing this problem. The analysis yields three important findings. To begin with, in the Kenyan political setting, manipulated information on TikTok manifests in various forms such as the doctored videos, fake news bulletins, disinformation campaigns organized, and hate speech promoted by algorithms against ethnic groups. Madung and Obilo (2021) showed that this kind of content regularly has millions of views before any moderation measure is implemented, which is indicative of both how fast the TikTok algorithm works and how slow its enforcement mechanisms are. The 2024 protests of Gen Z also highlighted how the state-funded disinformation networks use the platform to subvert civic movements (Amnesty International, 2025).

Second, TikTok uses various moderation tools: AI flagging, content labeling, banning users, and partnerships with fact-checkers. The tools can be considered a real effort to control political misinformation. Nevertheless, according to CIPESA (2022) and Gongane et al. (2022), these systems were trained mainly on English-language Western data and, every time, cannot identify Swahili-language and code-switched text. It is a structural issue, rather than a technical malfunction, and a sign of TikTok having under-invested in African moderation infrastructure. Third, TikTok has limited overall effectiveness in its moderation strategies in Kenya. The platform is reactive and not preventive. Bans are punitive actions that do not eradicate but push the disinformation creators away, as the study of MIS Quarterly (2026) confirms. More essentially, as Soliman (2024) claims, the business model of Tik Tok engagement is a structural tension, the very algorithms that promote engagement are the ones that, unintentionally, promote misleading content. To overcome the problem of political misinformation on Tik Tok in Kenya, solutions at the level of the platform are not enough. It will require context-sensitive AI systems that are trained on local language data, a real commitment to collaborating with Kenyan civil society organizations like Africa Check and Code for Africa, and more robust regulatory frameworks that will hold platforms accountable to principles of public interest as reflected by the Social Responsibility Media Theory. In the absence of these structural changes, the moderation of Tik Tok will be inadequate to safeguard the democratic communication space in Kenya.

REFERENCES

1. Aimeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation, and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(30). <https://doi.org/10.1007/s13278-023-01028-5>
2. Akmalia, W. (2019). *Divided Truth. Islamic Popularism and Electoral Political Disinformation*.
3. Ali, A., Bhattacharya, S., Ghosh, S., Ghosh, S., & Mukherjee, A. (2023). Countering Misinformation on social media through educational interventions: Evidence from a randomized experiment. *Journal of Development Economics*, 165, 103167. <https://doi.org/10.1016/j.jdeveco.2023.103167>
4. Almakaty, S. S. (2025). *Gatekeeper Theory in the Digital Media Age: A Comprehensive and Critical Literature Review*.
5. Asemah, E. S., Nwammuo, A. N., & Nkwam-Uwaoma, A. O. (2017). *Theories and models of Communication*.

6. Bowles, J. (2020). Countering misinformation via WhatsApp: Preliminary evidence from a field experiment in Brazil. *National Library of Medicine / PMC*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7357735/>
7. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
8. Capitol Technology University (2023). *TikTok and the War on Misinformation*. <https://www.captechu.edu/blog/tiktok-and-war-misinformation>
9. Castells, M. (2010). *The rise of the network society* (2nd ed.). Wiley-Blackwell.
10. Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford University Press.
11. Cobbe, J. (2021). *Algorithmic censorship by social platforms: Power and resistance*.
12. Collaboration on International ICT Policy for East and Southern Africa (CIPESA). (2022). *Disinformation in Kenya's political sphere: Actors, pathways and effects*. <https://cipesa.org/wp-content/files/Disinformation-in-Kenyas-Political-Sphere-Actors-Pathways-and-Effects.pdf>
13. Cookey, P. (2022). *Assessment of Citizen Journalists' Fake News in Media Space*.
14. Corsi, G., Iannelli, L., & Giglietto, F. (2024). Crowdsourcing the mitigation of disinformation and misinformation: The case of spontaneous community-based moderation on Reddit. *Computers in Human Behavior Reports*, 13, 100359. <https://doi.org/10.1016/j.chbr.2023.100359>
15. Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.
16. Creswell, J. W., & Miller, D. L. (2000). *Determining validity in qualitative inquiry*. *Theory Into Practice*, 39(3), 124–130. https://doi.org/10.1207/s15430421tip3903_2
17. Denzin, N. K., & Lincoln, Y. S. (2018). *The SAGE handbook of qualitative research* (5th ed.). SAGE Publications.
18. Ghauri, P., & Grønhaug, K. (2005). *Research methods in business studies: A practical guide* (3rd ed.). Pearson Education.
19. Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
20. Giuranno, M.G., & Manni, M.S. (2026). *The impact of Tiktok on elections: (Mis) information and Regulatory challenges*. <https://onlinelibrary.wiley.com/doi/10.1111/kykl.70054>
21. Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media: A review. *Social Network Analysis and Mining*, 12(1), 129. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9463759/>
22. Griffin, E., Ledbetter, A., & Sparks, G. (2022). *A first look at communication theory* (11th ed.). McGraw Hill.
23. Guest, G., Namey, E., & Mitchell, M. (2020). *Collecting qualitative data: A field manual for Applied research*. SAGE Publications.
24. Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Combating Fake News: Stakeholder Interventions and Potential Solutions. *IEEE Access*, 10. <https://doi.org/10.1109/ACCESS.2022.3193670>
25. Henn, M., Weinstein, M., & Foard, N. (2006). *A short introduction to social research*. SAGE Publications.
26. Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York University Press.
27. Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing* (2nd ed.). SAGE Publications.
28. Madrid-Morales, D., Tully, M., Mudavadi, K. C., Matanji, F., & Diop, L. (2025). Exploring audience agency in countering misinformation: An audience analysis in Kenya and South Africa. *International Journal of Communication*, 19, 2382–2406. <https://ijoc.org/index.php/ijoc/article/view/21936>
29. Matanji, F., Tully, M., Mudavadi, K. C., Diop, L., & Madrid-Morales, D. (2024). Media literacy and factchecking as proactive and reactive responses to misinformation in Kenya and Senegal. *African Journalism Studies*. <https://doi.org/10.1080/23743670.2024.2401782>
30. Matiko, S. (2025). *Tiktok top source of fake news in Kenya, Reuters report reveals*. <https://nation.africa/kenya/news/tiktok-top-source-of-fake-news-in-kenya-reutersreport-reveals-5091894#story>
31. Maweu, J. M. (2019). “Fake elections”? Cyber propaganda, disinformation and the 2017 general elections in Kenya. *African Journalism Studies*, 40(4), 62-76.

32. McKay, G. (2022). *Disinformation and democratic transition: A Kenyan case study*. Stimson Center. <https://www.stimson.org/2022/disinformation-and-democratic-transition-a-kenyan-case-study/>
33. Mudavadi, K. C., Tully, M., & Lomoywara, D. B. (2025). Exploring Kenyans' interactions with misinformation on WhatsApp. *Mobile Media & Communication*, 13(1), 69–87. <https://doi.org/10.1177/20501579241269651>
34. Nasebere, S.M. (2024). *Tiktok in Political Communication among Gen Z in Kenya: A study of 2022 General elections*.
35. Ndlela, N. (2020). Digital media, misinformation and political communication in Africa. *Journal of African Media Studies*, 12(2), 1–15.
36. Njagi, S. M. (2025). Disinformation as a threat to Kenya's National Security in the Contemporary Information Environment. *National Security: A Journal of the National Defence University-Kenya*, 3(2). <https://doi.org/10.64403/4efzpe03>
37. Obagwu, P. U., & Idris, K. M. (2019). Social responsibility theory of the press: a critique of its application and constraints. *International Journal of Innovative Research and Development*, 8(12), 26-35.
38. Okeya, O. & Asemah, E. S. (2021). *Politics of Media Ownership and Information Flow Interface in the Development of Nigeria*. In E. S. Asemah & K. Imafidon (Eds.). *Multidimensional Approaches to Sustainable Development in Nigeria*. Enugu: G.O. University Press.
39. Oloo Ong'ong'a, D. (2021). Countering the new media podia: Youth and "fake news" in Kenya. *Shanlax International Journal of Arts, Science and Humanities*, 8(4). <https://shanlaxjournals.in/journals/index.php/sijash/article/view/3303>
40. Rashidian, N., Brown, P. D., Hansen, E., Bell, E. J., & Albright, J. R. (2019). *Friend and foe: The platform press at the heart of journalism*.
41. Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, 291(3), 906-917.
42. Roberts, C. (2005). Gatekeeping theory: An evolution. *Recuperado el*, 16.
43. Roberts, J. (2026). The role of social media user: Affordances and social responsibility. *Communication and Democracy*, 1-18.
44. Roberts, S. T. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: journal of mobile media*, 10(1), 1-18.
45. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. In *ACM Transactions on Intelligent Systems and Technology* (Vol. 10, Number 3). <https://doi.org/10.1145/3305260>
46. Shirky, C. (2011). *The political power of social media: Technology, the public sphere, and political change*. *Foreign Affairs*, 90(1), 28–41.
47. Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6). <https://doi.org/10.1002/widm.1385>
48. Social media moderation and content generation: *Evidence from user bans*. (2026). *MIS Quarterly*, 50(1). <https://doi.org/10.25300/MISQ>
49. Social truth queries as a novel method for combating misinformation: Evidence from Kenya. (2024). *Journalism & Mass Communication Quarterly*, 101(2), 345–368. <https://doi.org/10.1177/10776990231215>
50. Soliman, W. (2024). Unethical but not illegal! A critical look at two-sided disinformation platforms. *Journal of Information Technology*, 39(1), 1–19. <https://doi.org/10.1177/02683962231157>
51. Suzor, N. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press.
52. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, 359(6380), 1146-1151.
53. Wang, X., Treen, K. M. D., & Lian, H. (2025). The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*, 6(1). <https://doi.org/10.37016/mr-2020-150>
54. Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.
55. Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.