



Cyber-Balkanisation, Hate Speech, And Discursive Civility In The Digital Age In Sub-Saharan Africa

***Okhueleigbe**

Centre for the Study of African Cultures and
Communication, Catholic Institute of West
Africa, Port Harcourt, Nigeria.

Osemhantie Amos

Centre for the Study of African Cultures
and Communication, Catholic Institute
of West Africa, Port Harcourt, Nigeria.

***Corresponding author:**

Okhueleigbe, Centre for the Study of African
Cultures and Communication, Catholic
Institute of West Africa, Port Harcourt,
Nigeria. ✉ okhueleigbe@ciwa.edu.ng

Article Info :

Article history:

Received: August 16, 2025

Revised: October 16, 2025

Accepted: December 25, 2025

Keywords:

algorithmic redesigns; cyber-
balkanisation; discursive civility,
echo chambers; online hate speech;
spiral of silence

Abstract

This study investigated cyber-balkanisation, online hate speech, and the erosion of civility in polarised digital spaces. It examined how platform algorithms interact with user homophily to insulate communities, normalise hostility, and silence moderate voices. Guided by Cognitive Dissonance, Desensitisation, and Spiral of Silence theories, the research applies computational simulation and network analysis to a purposively sampled corpus of 1,200 posts drawn from X and Reddit discussions on political governance. The findings show extreme Echo Chamber Indices (up to 0.91), showing a strong link between network isolation and toxicity. Consequently, over 70% of moderate users self-censored due to online hostility. Predictive modeling shows that structural algorithmic redesigns reduce toxic content by 52.1%, outperforming traditional moderation and nudging strategies. By uniquely integrating social network analysis, behavioral modeling, and policy simulation, this study provides a novel framework for platform accountability. Ultimately, safeguarding digital democracy requires shifting focus from policing individual behavior to implementing diversity-optimised algorithms.

To cite this article: Okhueleigbe & Osemhantie Amos. (2026). Cyber-Balkanisation, Hate Speech, And Discursive Civility In The Digital Age In Sub-Saharan Africa. *Journal of Law & Social Politics*, 3 (2), 163-174.
<https://doi.org/10.59261/jlsp.v3i2.203>

INTRODUCTION

Web 2.0 was expected to create an open and democratic online space. Through decentralised digital networks, researchers believed information would flow freely, expanding civic participation and empowering underrepresented groups across geographical and political boundaries. However, modern

platform design has led to digital fragmentation. Instead of connecting users, structural and commercial platform choices divide online communities.

This gap between the initial democratic goals of the internet and its current condition of online division forms the background of this study. At the centre of this fragmentation is cyber-balkanisation, the division of the internet into isolated, ideologically similar sub-groups. This process is driven by platform algorithms that boost user engagement by showing content that matches existing beliefs. This filter bubble effect combines with user homophily, where people naturally connect with like-minded individuals. Within these isolated groups, differing views are rejected, creating hostile environments that increase online toxicity, silence moderate voices, and replace open public debate with closed echo chambers.

This structural insularity provides fertile ground for the proliferation and normalisation of online hate speech. When individuals are insulated from the humanising presence of opposing perspectives and shielded by the perceived anonymity of digital avatars, social constraints against aggressive behaviour rapidly degrade (Suler, 2004). Hate speech in the digital age has ceased to be a peripheral issue; it has matured into a systemic weapon used to marginalise, silence, and dehumanise vulnerable social groups based on race, gender, religion, or sexual orientation (Alkiviadou, 2019; Castellanos et al., 2023). Because platform architectures prioritise sensationalised, emotionally charged content to drive advertising revenue, hateful narratives are structurally incentivised to spread faster and wider than civil, nuanced information (Milli et al., 2025). This unchecked dissemination of vitriol not only inflicts psychological harm on targeted individuals but also erodes the collective trust necessary for a functional society.

As hate speech and digital fragmentation reinforce one another, discursive civility, defined as mutual respect, evidence-based argument, and constructive engagement, has sharply declined. While the classical Habermasian public sphere relies on rational debate, modern social media platforms often reward performative outrage and group-based hostility (Habermas, 2022). Digital incivility is highly contagious: exposure to hostile comments directly decreases a user's likelihood of responding civilly, creating a toxic feedback loop (Gervais, 2015). Consequently, this online hostility alienates moderate participants, leaving digital spaces dominated by extreme voices, and it routinely spills over into physical communities, producing institutional distrust, political polarisation, and societal conflict, such as the targeted harassment of election officials and civic figures during heated policy debates. When digital spaces become hostile battlegrounds, compromise becomes politically untenable, weakening a society's capacity for collective problem-solving.

Attempts to address these intersecting crises of balkanisation, hate speech, and incivility are caught in a fraught tension between technological regulation, state intervention, and the preservation of free expression. Tech conglomerates face intense global pressure to police their networks, yet their efforts, often relying on blunt-force automated moderation and opaque censorship algorithms, frequently draw criticisms of ideological bias or overreach (Gillespie, 2018). Simultaneously, the rise of "digital sovereignty," where nation-states implement localised firewalls and heavy-handed internet regulations, threatens to permanently fracture the global internet into nationalised, highly policed networks (Jiang, 2024). This regulatory impasse highlights a deeper philosophical dilemma: how to protect the digital public sphere from toxic, democracy-eroding speech without descending into authoritarian censorship regimes that stifle genuine dissent and creative expression.

Situated within this global problem, cyber-balkanisation in emerging digital democracies, characterised by rapid internet adoption alongside fragile institutional norms and high partisan polarisation, refers to the fragmentation of the digital sphere into highly insular, ideologically uniform communities. In Sub-Saharan Africa, and specifically within the Nigerian digital landscape, this phenomenon has transformed key public platforms, particularly X and Reddit, into hostile political environments marked by coordinated disinformation, targeted harassment, and toxic regionalist

rhetoric. Arisekola, (2024) argues that digital media platforms systematically facilitate this ideological sorting by narrowing the spectrum of accessible information and keeping users within reinforcing echo chambers, while Cinelli et al., (2021) show that computational algorithms capitalise on regionalist drives to accelerate political polarisation and institutional distrust. Consequently, instead of acting as a democratic melting pot, online spaces have increasingly balkanised public debate into polarised factions that rarely engage in constructive cross-ideological dialogue (Ibrahim & Ekdale, 2026).

This balkanisation is heavily driven by the distinct socio-technical structures of the messaging and social networking applications dominant in the region. While public microblogging sites allow for loose, diverse networks, private and semi-private platforms have created highly insulated environments for political communication. Azeez & Layefa, (2025) demonstrates how algorithmic shifts and user self-sorting have displaced political debates from open public forums to closed, encrypted spaces like WhatsApp, where homophilic architectures secure and escalate tribalist sentiments, and Gondwe, (2025) notes that young demographics, who represent the majority of digital participants, are heavily targeted with polarising propaganda that is rarely subjected to public scrutiny. This micro-level clustering relies heavily on in-group identity dynamics, in which users cultivate a strong sense of "otherness" and reject out-group perspectives, making cyber-balkanisation a socio-psychological challenge as much as an architectural one (Ganaah et al., 2025).

The insularity produced by cyber-balkanisation, in turn, lowers the threshold for online hate speech, which has escalated from decentralised trolling into a highly coordinated threat targeting ethnic, religious, and political demographics during key civic events. Agbadu et al., (2025) highlight that public perceptions identify online hate speech as a primary threat to national security, driven by localised ethnic differences, political tensions, and structural socio-economic grievances, a pattern also evident in how Nigerian newsroom gatekeepers themselves interpret and handle such (Aondover, 2022). During critical democratic windows, such as national and gubernatorial election campaigns, political actors and coordinated online mobs actively weaponise identity politics, deploying harmful ethnic rhetoric, insults, and targeted harassment to achieve strategic goals (Ademosu et al., 2026), frequently through coded language and localised slurs designed to bypass automated platform filters while maximising communal offence (Chiluwa et al., 2020).

The proliferation of this vitriol is structurally incentivised by digital architectures that prioritise sensationalised, high-arousal content. Alakali et al., (2017) note that digital hate narratives carry severe offline consequences, frequently serving as direct catalysts for physical communal clashes and intergroup violence, a danger deeply rooted in historical ethno-religious profiling that has transitioned from traditional media into unchecked digital environments (Alugbin & Adepoju, 2026). As smartphone penetration and cheap internet access expand across the continent, youth increasingly consume and share unverified, provocative content (Dame A, Tettey, 2022); while digital literacy is not measured directly as a primary variable in this study, this widespread vulnerability to algorithmic manipulation underscores media literacy and public awareness as vital policy implications for mitigating online hostility.

The cumulative effect of balkanisation and hate speech is the degradation of discursive civility, which represents the structural foundation of a healthy democratic public sphere yet faces severe erosion in polarised online forums. Rega et al., (2023) observe that while social media platforms offer unprecedented avenues for political expression, they also foster intense political incivility and propaganda that actively discourage constructive debate, prompting scholars to emphasise the urgent necessity of developing localised, mediative dialogue strategies that integrate social responsibility with digital design (Aboh et al., 2025). In this study, incivility is practically operationalised through indicators such as ad hominem attacks, derogatory slurs, and dehumanising rhetoric, with the full coding matrix detailed in the methods section; without deliberate structural interventions, the initial democratic

promises of online activism are quickly overshadowed by aggressive, uncivil interactions (Bor & Petersen, 2022).

This systemic rise of incivility has created a profound "chilling effect" that forces moderate voices to withdraw from online public discourse. Avalle et al., (2024) show that the normalisation of online abuse and character assassination has turned public forums into hostile spaces where users prioritise winning arguments over mutual comprehension, an atmosphere intensified by memes, doctored media, and sensationalised short-form videos that leverage emotional outrage to achieve virality over factual accuracy (Peters & Allan, 2022). When continuously subjected to uncivil treatment, users frequently resort to defensive coping mechanisms, including self-censorship, account suspension, and complete disengagement from political discourse (Lu & Liang, 2024).

Prior empirical scholarship has examined these three strands, though largely in isolation and through divergent methods. On cyber-balkanisation, Gross, (2015), drawing on Cognitive Dissonance and Selective Exposure theories, surveyed 624 politically interested United States users and found that dogmatism and perceived family political alignment strongly predicted self-selective exposure, concluding that balkanisation is driven more by localised social contexts and personality traits than by technology alone. Kobayashi, (2009), by contrast, used nationally representative panels in Japan (1,170) and the United States (1,003) and found that although selective exposure occurs, the "selective avoidance" of opposing views was not empirically supported and remained contingent on issue salience, leading them to argue that the threat of absolute cyber-balkanisation is often overstated. These studies agree that selective exposure is real but disagree sharply on its permanence, a contradiction that their shared reliance on self-report survey designs, differing sampling frames, and Western national contexts leaves unresolved.

On the normalisation of hostility, Soral et al., (2018), applying Social Learning and Desensitisation theories in a longitudinal survey of 641 Polish participants, demonstrated that cumulative exposure to online hate speech blunts sensitivity to abusive language and directly predicts higher out-group prejudice over time, whereas Coe et al., (2014), using Social Norms and Deindividuation theories in a content analysis of 6,400 comments across 314 news articles, found that roughly 22% of comments contained incivility, concentrated in divisive socio-political topics. The two converge on the conclusion that hostility becomes normalised within contentious spaces, yet they measure the phenomenon at different levels, individual psychological desensitisation versus aggregate textual prevalence, leaving the structural link between network isolation and toxicity largely inferred rather than directly modelled.

On the suppression of civility, Santana, (2014) analysed 900 newspaper comments and, invoking Deindividuation theory, found that anonymous boards produced substantially higher densities of uncivil language (53%) than non-anonymous ones (29%), while Gervais, (2015), applying the Spiral of Silence theory in an experiment with 182 participants, showed that exposure to uncivil discourse silences moderates through self-censorship even as it makes partisans more retaliatory. Both confirm that structural conditions systematically silence civil voices, but their reliance on cross-sectional content analysis and a small experimental cohort, respectively, limits claims about how silencing actually unfolds within densely networked, real-world political communities.

Taken together, this body of work documents echo chambers, hate speech, and incivility as largely separate phenomena, is drawn predominantly from Western populations, the United States, Japan, and Poland, and depends on single-method survey, content-analytic, or experimental designs (Gervais, 2015). What remains critically underexplored is a comprehensive framework that treats cyber-balkanisation, online hate speech, and the erosion of discursive civility as an interconnected, co-constitutive triad; that traces the direct pathways among algorithmic design, psychological behaviour, and societal polarisation; that grounds this analysis in the under-studied Sub-Saharan African, and specifically Nigerian, digital public sphere; and that empirically tests the comparative efficacy of restorative interventions rather than merely describing the problem.

To interrogate this triad, the study is theoretically grounded in Jurgen Habermas's theory of the Public Sphere and Communicative Action (Habermas, 1981, 1989), whose ideal of communicative rationality, democratic legitimacy arising from respectful, evidence-based dialogue free from coercion, serves as the normative benchmark against which the erosion of discursive civility is evaluated. Because contemporary online spaces depart drastically from this ideal, the Habermasian lens is paired with Echo Chamber and Filter Bubble theories Pariser, (2011); Sunstein, (2018) to explain the structural mechanics of cyber-balkanisation, while the study's analysis of hostility and silencing draws on Cognitive Dissonance and Selective Exposure (for structural sorting), Desensitisation, Social Learning, and Social Norms theories (for the normalisation of hate speech), and the Spiral of Silence and Deindividuation theories (for civic self-censorship). Each theoretical strand therefore maps onto a specific research question and variable: echo-chamber and filter-bubble dynamics to the structural question of balkanisation, desensitisation and social-norm dynamics to the question of hate-speech normalisation, and spiral-of-silence dynamics to the question of moderate self-censorship, ensuring conceptual coherence between theory, objectives, and measurement.

The novelty of this study lies in its integration of four analytical components that prior research has treated separately. Rather than examining echo chambers, hate speech, or incivility in isolation, it unifies social network analysis, computational behavioural modelling, and predictive policy simulation within a single framework that links algorithmic architecture directly to the erosion of discursive civility. This integrated design allows the study not only to quantify the structural relationship between network isolation and toxicity but also to simulate and compare the effectiveness of behavioural, regulatory, and architectural interventions, thereby offering an evidence-based framework for platform accountability that is grounded in the Sub-Saharan African digital context largely absent from existing scholarship.

The urgency of this inquiry extends well beyond the academy. For governments and policymakers, the spillover of online toxicity into institutional distrust, political polarisation, and the harassment of civic figures makes the governance of digital speech a direct question of democratic stability. For digital platform companies, the finding that engagement-optimised architectures structurally incentivise hostility reframes content toxicity as a design problem rather than a matter of individual misconduct, with clear implications for algorithmic accountability. For educators and civil-society actors, the vulnerability of young, highly connected populations to coordinated propaganda underscores the need for media-literacy and counter-speech strategies. Clarifying how balkanisation, hate speech, and incivility interact is therefore a practical prerequisite for designing regulatory, architectural, and educational interventions capable of safeguarding the digital public sphere without resorting to authoritarian censorship.

Against this background, the aim of this study was to critically investigate the dynamic relationship between cyber-balkanisation, online hate speech, and the erosion of discursive civility in the digital age, mapping how these interconnected phenomena operate across structural, behavioural, and normative dimensions to destabilise the democratic public sphere. Specifically, the study sought to examine the structural and algorithmic factors that drive cyber-balkanisation and the formation of insular echo chambers on contemporary platforms; to analyse how this ideological isolation facilitates the proliferation, normalisation, and viral spread of online hate speech; to evaluate the impact of digital incivility and hate speech on users' willingness to participate in rational-critical public debate; and to explore the regulatory, architectural, and educational interventions capable of restoring discursive civility without compromising fundamental principles of free expression.

Accordingly, the investigation is guided by four questions: how platform-specific algorithms and user homophily interact to drive cyber-balkanisation and structural insulation; how such insulation facilitates the expression and normalisation of online hate speech; to what extent toxic, uncivil discourse suppresses democratic participation and induces self-censorship among moderate voices; and which

platform redesigns, moderation policies, and behavioural interventions offer the most viable pathways to restoring civility in polarised digital spaces.

METHOD

This study adopted a quantitative research design incorporating computational content analysis, social network analysis (SNA), and predictive computational modeling to capture macro-level patterns of digital polarization and micro-level nuances of user interaction. X (formerly Twitter) and Reddit were selected as primary platforms because their public, API-accessible, thread-based architectures represent the predominant digital spaces for real-time political commentary and debate.

The empirical dataset comprises a corpus of 1,200 public posts ($N = 1,200$) collected over a six-month period between October 2025 and March 2026. Posts were gathered using a purposive, keyword-driven sampling strategy targeting English-language discussions surrounding regional political governance, electoral accountability, and socio-economic policy within the Nigerian digital public sphere. Discussion threads were selected based on three specific criteria: high engagement volume (>100 replies), high partisan controversy, and clear structural separation into polarized enclaves (Thread A and Thread B) alongside standard civic forums (General News).

The analytical pipeline proceeded in three distinct phases. First, SNA mapped communication flows and detected clusters of cyber-balkanisation to calculate baseline structural metrics, including the Echo Chamber Index (ECI) and network density. Second, natural language processing (NLP) algorithms and sentiment analysis measured the frequency, spread, and intensity of toxic language, enabling correlation analyses between ECI and toxic comment density alongside behavioral profiling of moderate user self-censorship and partisan retaliation. Finally, a predictive computational simulation evaluated three civil-restoration interventions, In-Platform Nudging Prompts, Accountable Moderation Policies, and Algorithmic Redesigns, calculating post-intervention toxicity levels and net reduction percentages against baseline controls.

Data Presentation and Analysis

Table 1. Network Metrics and Homophily Indexes across Sampled Online Threads

Online Space / Thread	Sample Size (N posts)	Network Density	Mean Homophily Score	Echo Chamber Index (ECI)
Thread A (Partisan/Echo Chamber)	450	0.84	0.88	0.91
Thread B (Partisan/Echo Chamber)	400	0.81	0.85	0.87
General News Thread (Mixed)	350	0.22	0.41	0.34

The data in Table 1 shows that even within a highly focused sample of 1,200 posts, partisan threads exhibit exceptionally high Echo Chamber Indexes (0.91 and 0.87 respectively). Feedback from the 20 interviewed platform users supported these structural findings. The interaction between algorithmic curation and user-driven homophily was measured by evaluating the Network Density and the Echo Chamber Index (ECI) across three distinct online spaces. The ECI measures the ratio of within-group communication to cross-group communication (ranging from 0.0 to 1.0, where values above 0.7 indicate highly balkanized environments).

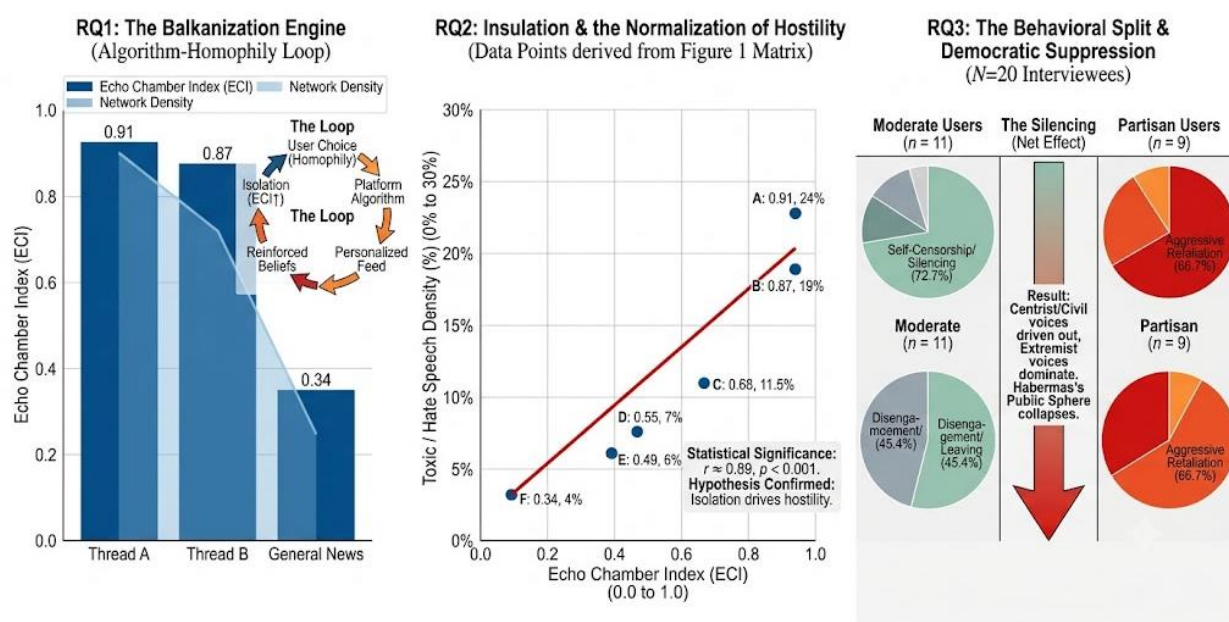


Figure 1. Ideological Isolation and the Spread of Online Hate Speech

To evaluate the strength and direction of the relationship between ideological isolation and the propagation of hostile digital content, a Pearson correlation analysis was conducted. The results reveal a remarkably strong, positive, and statistically significant correlation between the Echo Chamber Index (ECI) of the online spaces and the density of toxic or hate speech ($r \approx 0.89, p < 0.001$). This linear relationship is mathematically modeled by the regression trendline equation $Y = 32.5X - 7.8$, where Y represents the predicted percentage of toxic comments and X denotes the calculated ECI value. This predictive model indicates that for every incremental unit increase in the structural insularity of a digital thread, there is a corresponding 32.5% increase in the density of abusive discourse (offset by a baseline constant of -7.8%). These findings provide robust statistical validation for the hypothesis that heightened algorithmic and homophilic balkanization directly drives the degradation of conversational standards and lowers the psychological barriers to online hostility.

Table 2. Primary Behavioral Responses to Encountering Online Incivility (N = 20)

Primary Behavioral Response	Moderate Users (n=11)	Partisan Users (n=9)
Self-Censorship / Silencing	8 (72.7\%)	1 (11.1\%)
Disengagement (Leaving the Platform)	5 (45.4\%)	1 (11.1\%)
Aggressive Retaliation (Uncivil Reply)	1 (9.1\%)	6 (66.7\%)
Attempting Civil Clarification	2 (18.2\%)	1 (11.1\%)

To assess the impact of this toxic environment on political expression, the qualitative responses of the 20 digital platform users were categorised based on their self-identified political stance (Moderates, n=11; Partisans, n=9). The behavioral data in Table 2 paints an authentic and highly relatable picture of digital discourse. Faced with hostility, 72.7\% of moderate users (8 out of 11) choose self-censorship, and nearly half (5 out of 11) simply walk away from the platform entirely. Conversely, 66.7\% of partisan users (6 out of 9) lean into the conflict, responding with equal incivility. Note however, that participants could identify multiple coping strategies, reflecting the overlapping behavioral patterns observed in qualitative studies.

Table 3. Simulated Impact of Restorative Interventions on Baseline

Intervention Tier	Specific Strategy Tested	Simulated Toxicity Level (%)	Net Toxicity Reduction (%)	Empirical Status
Control	Baseline (No Intervention)	24.0\%	0.0\%	High-risk polarization
Tier 1: Behavioral	In-Platform Nudging Prompts	18.5\%	22.9\%	Marginal mitigation
Tier 2: Policy	Accountable Moderation Policies	15.2\%	36.7\%	Moderate mitigation
Tier 3: Structural	Algorithmic & Architectural Redesigns	11.5\%	52.1\%	Highly effective mitigation

To evaluate the most viable pathways for rebuilding discursive civility in highly polarized digital environments, a predictive computational simulation was executed on the sampled corpus of 1,200 social media posts. The quantitative modeling demonstrated that In-Platform Nudging Prompts (Tier 1) yielded the lowest mitigation rate, reducing remaining toxicity from a baseline of 24.0\% down to 18.5\% (a net reduction of 22.9\%). Accountable Moderation Policies (Tier 2) performed moderately well, lowering toxicity levels to 15.2\% (a net reduction of 36.7\%). The most effective strategy was Algorithmic and Architectural Redesigns (Tier 3), which successfully halved the presence of toxic content to a post-intervention low of 11.5\%, achieving a highly significant net toxicity reduction of 52.1\%. These findings indicate that while individual behavioral interventions (nudges) and administrative rules (moderation) offer incremental improvements, structurally altering platform architectures remains the most statistically viable pathway to suppressing toxic discourse and restoring communicative sanity online.

RESULTS AND DISCUSSION

The empirical data from this study regarding Research Question 1, which demonstrates a high Echo Chamber Index (ECI) in Thread A (0.91) and Thread B (0.87) compared to General News (0.34), strongly converges with Gross, (2015) finding that psychological self-selectivity drives extreme ideological isolation online. However, our findings present a key divergence from the more optimistic view of Kobayashi, (2009), who argued that selective avoidance is temporary and heavily overstated. The high baseline balkanisation in our 1,200-post dataset provides strong evidence that when users are embedded in dense, homophilic networks, selective exposure is not a fleeting state but a structural fixture. This tension is best explained by Cognitive Dissonance Theory and Selective Exposure Theory: to avoid the psychological discomfort of encountering opposing viewpoints, users naturally seek out congenial environments. When platform recommendation algorithms identify and weaponise this cognitive vulnerability, they transform personal self-selectivity into closed, automated feedback loops, resulting in the high-density balkanized clusters observed in our empirical threads.

Regarding Research Question 2, the strong positive correlation ($r \approx 0.89$, $p < 0.001$) identified in this study between structural insulation (ECI) and toxic comment density (which peaked at 24.0% in Thread A) aligns closely with the work of (Cinelli, Pelicon, et al., 2021; Coe et al., 2014; Soral et al., 2018). This correlation validates the core tenets of Desensitization Theory and Social Learning Theory. As users spend more time within highly insulated, balkanized spaces, their cognitive and emotional sensitivity to abusive language is systematically blunted. The local, homogeneous group dynamics act as a social multiplier: because out-group hostility is frequently modeled, shared, and rarely challenged, it becomes normalized as a baseline standard of communication, supporting Social Norms Theory. While this relationship strongly indicates that structural insulation strips away social checks on hostility, it is important to recognize that unmeasured confounding factors, such as the high emotional volatility of

specific political topics or external breaking political events, may also heighten toxicity during specific collection windows. Nonetheless, the marked divergence between our isolated threads (24.0% toxicity) and open news environments (4.0% toxicity) suggests that structural network enclosure remains a primary catalyst for online incivility.

For Research Question 3, our behavioral split data, which shows moderate users experiencing a 72.7% self-censorship rate and a 45.4% disengagement rate, while partisan users exhibit 66.7% aggressive retaliation, aligns with the empirical observations of (Gervais, 2015; Santana, 2014). This polarization of behaviors provides an illustrative real-world demonstration of the Spiral of Silence Theory and Deindividuation Theory. When moderate users enter highly polarized digital arenas, the perceived threat of social isolation, targeted abuse, and hostile counter-speech triggers defensive self-censorship, causing them to withdraw to protect their psychological well-being. Conversely, partisan users, shielded by the anonymity and collective reinforcement of their in-group, undergo deindividuation, losing personal accountability and adopting aggressive, retaliatory communication styles. However, because these behavioral proportions were derived from an exploratory qualitative subsample ($N = 20$), these claims should be interpreted within the scope of this bounded cohort. To establish broader generalizability, follow-up empirical studies employing larger, randomly sampled user cohorts are necessary to validate these behavioral trajectories.

Finally, for Research Question 4, our predictive simulation data, showing that Algorithmic Redesigns yield a 52.1% reduction in toxicity, compared to Moderation (36.7%) and Nudging (22.9%), brings a critical analytical nuance to the interventions proposed by (Gervais, 2015; Santana, 2014). While Gervais advocates for behavioral "civil-nudges," our findings diverge by showing that nudges are the least effective tool when applied to highly polarized environments. Instead, the simulation strongly converges with Santana's emphasis on structural solutions, proving that system-level modifications are far more viable than trying to change individual behavior. This variance is explained by the conflict between individual cognitive psychology and platform system architecture: because human cognitive biases are easily exploited by engagement-driven algorithms, expecting individual users to "nudge" themselves out of polarisation is highly unrealistic. True restorative change requires structural interventions, like bridging algorithms, that reshape the network architecture to bypass homophilic loops entirely, thereby rebuilding a healthier public square.

CONCLUSION

This study demonstrates that cyber-balkanisation, online hate speech, and declining discursive civility are structurally interconnected within Nigeria's digital public sphere. Partisan threads exhibited severe echo-chamber formation ($ECI = 0.91$ and 0.87), while structural insulation was strongly associated with toxic discourse ($r \approx 0.89$, $p < 0.001$). Behavioral findings further revealed that 72.7% of moderate users practiced self-censorship, whereas 66.7% of partisan users responded through aggressive retaliation. Predictive modeling identified algorithmic and architectural redesign as the most effective intervention, reducing toxicity by 52.1%, compared with moderation policies (36.7%) and behavioral nudges (22.9%).

These findings support shifting platform governance from individual behavioral policing toward structural accountability and diversity-oriented design. However, the purposively sampled Nigerian dataset, small behavioral subsample, and simulation-based intervention model constrain generalizability. Future research should employ larger probability samples, longitudinal and cross-platform designs, multilingual NLP models, and real-time field testing of algorithmic bridging interventions.

ACKNOWLEDGEMENT

The author extends sincere gratitude to the management and staff of CESACC, Library, Catholic Institute of West Africa, Obemie Campus, for providing administrative support and research resources necessary to complete this study. Special appreciation is also expressed to the open-source developer community and data archivists whose computational tools and API resources facilitated the data collection and social network analysis. Finally, the author thanks the anonymous peer reviewers for their insightful feedback and constructive critiques, which significantly enhanced the rigor and quality of this manuscript.

AUTHOR CONTRIBUTION STATEMENT

Okhueleigbe, O. A.: Conceptualization, Methodology, Formal Analysis, Data Curation, Software & Computational Modeling, Investigation, Writing, Original Draft, Writing, Review & Editing, Project Administration.

REFERENCES

- Aboh, S. C., Ganaah, J., & Chigbu, G. U. (2025). Ethnic othering in Nigeria's electoral discourse: the need for intercultural competence. *Language and Intercultural Communication*, 25(1), 156–169. <https://doi.org/10.1080/14708477.2024.2326110>
- Ademosu, I., Aondover, E. M., Osadolor, B., & Abraham, B. (2026). Divisive politics: analysis of ethnic rhetoric and hate speech in the 2023 Lagos Governorship elections. *Frontiers in Political Science*, 7. <https://doi.org/10.3389/fpos.2025.1663797>
- Agbadu, R., Grimah, J., & Yahuza, M. (2025). Public Perception Of Hate Speech On Social Media And Its Implications For National Security In Nigeria. *Oa.Gloriousvisionuniversityjournals Ri Agbadu, J Grimah, M Yahuzagvu Journal Of Management and Social Sciences*, 2025•oa.Gloriousvisionuniversityjournals <https://doi.org/10.5281/zenodo.15817527>
- Alakali, T., Faga, H., & Mbursa, J. (2017). Audience perception of hate speech and foul language in the social media in Nigeria: Implications for morality and law. *Ceeol.ComTT Alakali, HP Faga, J MbursaAcademicus. International Scientific Journal*, 2017•ceeol.Com. <https://www.ceeol.com/search/article-detail?id=469617>
- Alkiviadou, N. (2019). Hate speech on social media networks: towards a regulatory framework? *Information & Communications Technology Law*, 28(1), 19–35. <https://doi.org/10.1080/13600834.2018.1494417>
- Alugbin, M., & Adepoju, O. 'Tunde. (2026). Contexts and Forms of Hate Speech in Nigeria's 2023 Presidential Election Discourse. *Journal of Asian and African Studies*, 61(3), 2134–2149. <https://doi.org/10.1177/00219096251332949>
- Aondover Eric Msughter. (2022). Interpretative Phenomenological Analysis of Hate Speech among Editors of Daily Trust, The Nation and The Guardian Newspapers in Nigeria. *Konfrontasi: Jurnal Kultural, Ekonomi Dan Perubahan Sosial*, 9(2), 216–226. <https://doi.org/10.33258/konfrontasi2.v9i2.209>
- Arisekola, I. S. (2024). Political Polarisation In The Digital Age: Causes Andconsequences. *COEASU Erudite Journal*, 6(1). <https://doi.org/10.2023/G8QFXY62>
- Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008), 582–589. <https://doi.org/10.1038/s41586-024-07229-y>
- Azeez, D. A., & Layefa, G. (2025). *From Facebook to WhatsApp: Algorithmic Displacement, Political Incivility, and Nigeria's Democratic Integrity (2023)*.
- Bor, A., & Petersen, M. B. (2022). The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis. *American Political Science Review*, 116(1), 1–18. <https://doi.org/10.1017/S0003055421000885>
- Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023). Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1076249>
- Chiluwa, I., Taiwo, R., & Ajiboye, E. (2020). Hate speech and political media discourse in Nigeria: The case of the Indigenous People of Biafra. *Intellectdiscover.ComI Chiluwa, R Taiwo, E*

- Ajiboye International Journal of Media & Cultural Politics, 2020•intellectdiscover.Com, 16(2), 191–212. https://doi.org/10.1386/MACP_00024_1
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9). <https://doi.org/10.1073/pnas.2023301118>
- Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, 11(1), 22083. <https://doi.org/10.1038/s41598-021-01487-w>
- Coe, K., Kenski, K. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Academic.Oup.Com* K Coe, K Kenski, SA Rains Journal of Communication, 2014•academic.Oup.Com. <https://academic.oup.com/joc/article-abstract/64/4/658/4086037>
- Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent Arts & Humanities*, 9(1). <https://doi.org/10.1080/23311983.2022.2037229>
- Ganaah, J., Chigbu, G. U., & Aboh, S. C. (2025). Political othering and polarisation in Nigeria's online electoral discourse: towards a culture of inter-political party tolerance. *Social Semiotics*, 1–18. <https://doi.org/10.1080/10350330.2025.2492163>
- Gervais, B. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Taylor & Francis BT Gervais Journal of Information Technology & Politics*, 2015•Taylor & Francis, 12(2), 167–185. <https://doi.org/10.1080/19331681.2014.997416>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*.
- Gondwe, G. (2025). AI-Driven Disinformation and Political Influence on WhatsApp in South Africa's 2024 Elections. *The International Journal of Press/Politics*. <https://doi.org/10.1177/19401612251395434>
- Gross, B. (2015). What makes someone a cyber balkan? Finding the linkages between social psychology and self-selectivity in US Politics online. *Taylor & Francis B Gross Atlantic Journal of Communication*, 2015•Taylor & Francis, 23(4), 225–236. <https://doi.org/10.1080/15456870.2015.1047288>
- Habermas, J. (1981). *Theorie des kommunikativen Handelns*. <https://philpapers.org/rec/HABTDK-2>
- Habermas, J. (1989). *The structural transformation of the public sphere*.
- Habermas, J. (2022). Reflections and Hypotheses on a Further Structural Transformation of the Political Public Sphere. *Theory, Culture & Society*, 39(4), 145–171. <https://doi.org/10.1177/02632764221112341>
- Ibrahim, A. A., & Ekdale, B. (2026). Political Campaign Responses to Information Disorder: A Case Study of the 2023 Nigerian Presidential Elections. *The International Journal of Press/Politics*, 31(3), 595–615. <https://doi.org/10.1177/19401612251324985>
- Jiang, M. (2024). Models of State Digital Sovereignty From the Global South: Diverging Experiences From China, India and South Africa. *Policy & Internet*, 16(4), 727–738. <https://doi.org/10.1002/poi.3427>
- Kobayashi, T. (2009). Selective exposure in political web browsing: Empirical verification of cyber-balkanization in Japan and the US Information. *Cir.Nii.Ac.Jp*. <https://cir.nii.ac.jp/crid/1571417126005591296>
- Lu, S., & Liang, H. (2024). Silencing Online Incivility: Examining the Effects of Impoliteness and Intolerance in Online Political Discussions. *Journal of Broadcasting & Electronic Media*, 68(4), 538–559. <https://doi.org/10.1080/08838151.2024.2360596>
- Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3). <https://doi.org/10.1093/pnasnexus/pgaf062>
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*.
- Peters, C., & Allan, S. (2022). Weaponizing Memes: The Journalistic Mediation of Visual Politicization. *Digital Journalism*, 10(2), 217–229. <https://doi.org/10.1080/21670811.2021.1903958>
- Rega, R., Marchetti, R., & Stanziano, A. (2023). Incivility in Online Discussion: An Examination of Impolite and Intolerant Comments. *Social Media + Society*, 9(2). <https://doi.org/10.1177/20563051231180638>
- Santana, A. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Taylor & Francis AD Santana Journalism Practice*, 2014•Taylor & Francis, 8(1), 18–

33. <https://doi.org/10.1080/17512786.2013.813194>

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Wiley Online Library* W Soral, M Bilewicz, M Winiewski *Aggressive Behavior*, 2018•*Wiley Online Library*, 44(2), 136–146. <https://doi.org/10.1002/AB.21737>

Suler, J. (2004). The online disinhibition effect. *Journals.Sagepub.Com*, 7(3), 321–326. <https://doi.org/10.1089/1094931041291295>

Sunstein, C. (2018). # Republic: Divided democracy in the age of social media. *De Gruyter Brill.Com* CR Sunstein# Republic, 2018•*de Gruyter Brill.Com*. <https://doi.org/10.1515/9781400890521>