



Fake Accounts and Political Manipulation: Topic and Sentiment Analysis in Indonesia

Cuentas falsas y manipulación política: análisis temático y de sentimientos en Indonesia

Rendra Widyatama*, Universitas Ahmad Dahlan, (Indonesia) (rendrawidyatama@fsbk.uad.ac.id) (<https://orcid.org/0000-0003-0448-3399>)
Imam Adi Nata, Universitas Tidar, Yogyakarta (Indonesia) (imamadinata@untidar.ac.id) (<https://orcid.org/0009-0005-6857-9981>)
Muhammad Thoyib Amali, Universitas Ahmad Dahlan, Yogyakarta, (Indonesia) (muhammad.amali@comm.uad.ac.id) (<https://orcid.org/0009-0008-7835-9239>)
Endah Sari Manganti, Universitas Ahmad Dahlan, Yogyakarta (Indonesia) (2300030245@webmail.uad.ac.id) (<https://orcid.org/0009-0009-6209-4392>)
Olivia Istifany, Universitas Ahmad Dahlan, Yogyakarta (Indonesia) (2300030240@webmail.uad.ac.id) (<https://orcid.org/0009-0004-5415-9196>)

* Indicates the corresponding author

ABSTRACT

Fake accounts have evolved from mere tools of online deception into strategic infrastructures of political manipulation that systematically shape public opinion and intensify polarization. This study aims to uncover how fake accounts operate as coordinated actors that exploit platform architectures to amplify divisive narratives in Indonesia's digital public sphere. Using a Natural Language Processing (NLP) pipeline—data collection, preprocessing, topic modeling, sentiment analysis, and interpretive mapping—this research analyzed 8,102 posts generated by 100 identified fake accounts across X (Twitter) and Instagram. The findings reveal that negative, morally charged narratives—although constituting a small fraction of all content—disproportionately drive engagement and virality, particularly on X, whose open architecture facilitates mass diffusion. In contrast, Instagram's private network design confines manipulative content within insulated circles, suggesting differentiated strategies of disinformation. The study contributes new empirical evidence on how inauthentic actors leverage platform logics to influence collective sentiment. Beyond theoretical implications, it recommends multi-level interventions through digital literacy education, platform governance reforms, and regulatory oversight to mitigate the systemic risks of coordinated inauthentic behavior.

RESUMEN

Las cuentas falsas han evolucionado de ser meras herramientas de engaño en línea a convertirse en infraestructuras estratégicas de manipulación política que moldean sistemáticamente la opinión pública e intensifican la polarización. Este estudio busca revelar cómo dichas cuentas operan como actores coordinados que aprovechan la arquitectura de las plataformas para amplificar narrativas divisivas en la esfera pública digital de Indonesia. Mediante un flujo de trabajo de procesamiento de lenguaje natural (PLN) —que abarca la recopilación de datos, el preprocesamiento, el modelado de temas, el análisis de sentimiento y el mapeo interpretativo—, la investigación analizó 8.102 publicaciones generadas por 100 cuentas falsas identificadas en X (Twitter) e Instagram. Los hallazgos revelan que las narrativas negativas y cargadas de contenido moral —aunque representan una pequeña fracción del contenido total— impulsan de manera desproporcionada la interacción y la viralidad, especialmente en X, cuya arquitectura abierta facilita la difusión masiva. En cambio, el diseño de red privada de Instagram confina el contenido manipulador a círculos cerrados, lo que sugiere la existencia de estrategias de desinformación diferenciadas. El estudio aporta nueva evidencia empírica sobre cómo los actores no auténticos se valen de la lógica de las plataformas para influir en el sentimiento

colectivo. Más allá de las implicaciones teóricas, se recomiendan intervenciones multinivel —mediante educación en alfabetización digital, reformas en la gobernanza de las plataformas y supervisión regulatoria— para mitigar los riesgos sistémicos asociados al comportamiento no auténtico coordinado.

KEYWORDS | PALABRAS CLAVE

Fake Accounts, Political Manipulation, Disinformation, Sentiment Analysis, Digital Polarization, Indonesia, Platform Governance.

Cuentas falsas, desinformación, polarización, elecciones, indonesia, redes sociales.

1. Introduction

The proliferation of fake accounts on social media has become a major challenge for contemporary democracy. These accounts frequently disseminate disinformation (Rocha et al., 2023), manipulate public opinion (Bovet & Makse, 2019), and deepen societal polarization (Törnberg, 2022). Empirical evidence shows that coordinated activities by fake accounts—including bots and trolls—can influence electoral outcomes and erode trust in democratic institutions through the mass diffusion of misleading narratives (Coffé, 2017; Widyatama & Mahbob, 2024). Globally, disinformation campaigns driven by fake accounts have been documented in countries such as the United States, Brazil, and India, highlighting their threat to public deliberation and democratic resilience (Guess et al., 2020). Academic findings further demonstrate the intensification of digital astroturfing and coordinated campaigns using fake accounts in the run-up to elections (Rumata & Nugraha, 2020; Yerlikaya & Aslan, 2020).

However, despite increasing attention to the phenomenon, previous studies have tended to treat fake accounts merely as vehicles of disinformation, rather than as *strategic infrastructures* that deliberately exploit platform architectures to engineer public sentiment and influence policy discourse. This distinction is crucial: fake accounts do not only spread falsehoods but also manipulate emotional engagement, algorithmic visibility, and cross-platform virality—turning technical affordances into instruments of political control.

In Indonesia, the issue of fake accounts is particularly salient given the central role of social media in political discourse (Lim, 2017; Santoso et al., 2020). Two of the most prominent platforms, X (formerly Twitter) and Instagram, have become key arenas for debates over elections, governance, and identity politics. Political communication on these platforms intensifies in the pre-election period, providing fertile ground for fake accounts to disseminate manipulative and provocative content. Astroturfing—defined as the creation of a false perception of grassroots support through covertly organized activities (Zerback et al., 2021)—is a recurring tactic. Yet, systematic studies examining the content and sentiment of messages spread by these accounts remain scarce.

The Indonesian digital ecosystem offers a unique context where media logic, political mobilization, and digital literacy intersect. Despite numerous digital education programs and governance reforms, fake accounts continue to thrive—indicating a persistent gap between regulatory intent and algorithmic realities. This paradox raises an important question: why do interventions fail to reduce manipulation, and how do platform architectures enable such resilience?

Empirical research adopting a computational social science perspective on the role of fake accounts in shaping political narratives in Indonesia is still limited. While previous studies have explored disinformation and online polarization, the specific contribution of fake accounts—particularly in the distribution of topics and sentiment orientations—has been insufficiently addressed. A bibliometric search in Dimensions.ai over the past five years, using keywords such as “fake accounts,” “social media,” “topic distribution,” “sentiment orientation,” and “Indonesia,” identified only five relevant articles, as summarized in Table 1.

Table 1: International Journal Publications on Fake Accounts in Indonesia.

Authors (Year)	Methods and Findings
Subekti et al. (2025)	Qualitative case study on candidate-led disinformation. Disinformation was most prevalent as video and text on Facebook, YouTube, and TikTok.
Rumata and Nugraha (2020)	Qualitative case study of disinformation flows on Facebook, YouTube, and TikTok. Video/text were the primary carriers, significantly shaping electoral dynamics.
Wendratama et al. (2023)	Qualitative in-depth interviews (five cities: Jakarta, Banda Aceh, Yogyakarta, Makassar, Jayapura) on how women used WhatsApp groups and handled harmful content during the election.
Mohamad et al. (2025)	Highlights links between fake news and political polarization during the presidential campaign across X, Facebook, Instagram; Twitter/X stood out as the key channel for false narratives.
Bajari et al. (2023)	Virtual ethnography of identity-based hate speech on X; hate-laden comments/posts consistently triggered two-way and multi-directional interaction patterns.

This scarcity of empirical studies underscores a clear research gap—both theoretically and practically. Theoretically, the lack of computational mapping of fake account narratives limits our understanding of how digital architectures structure manipulative discourse. Practically, without this understanding, policy-makers, educators, and platform regulators lack evidence-based insights to design preventive measures.

Therefore, this study aims to: (1) Analyze the dominant political narratives produced by fake accounts

in Indonesia across two major platforms (X and Instagram); (2) Examine the sentiment orientation of these narratives to identify how emotional tone affects engagement and diffusion; and (3) Interpret how platform architecture mediates the visibility and impact of manipulative content.

By applying Natural Language Processing (NLP) techniques—including topic modeling and sentiment analysis—this study provides a computational framework to understand how coordinated inauthentic actors exploit different platform logics. In doing so, it contributes to three key domains: (a) communication and media studies, by modeling narrative manipulation through data-driven analysis; (b) political communication, by revealing how digital infrastructures amplify polarization; and (c) digital education and policy, by offering evidence-based recommendations for platform governance and literacy interventions.

Ultimately, this research not only fills a theoretical and methodological gap but also aligns with broader social needs: strengthening democratic resilience through digital literacy and promoting regulatory frameworks that can anticipate the evolving tactics of inauthentic online behavior.

2. Material and Method

This study employed a computational approach based on Natural Language Processing (NLP) to analyze conversations generated by fake accounts on social media. This method enables the processing of large-scale data and facilitates thematic issue mapping (Sharma et al., 2023). The research consisted of six stages: data collection, data preprocessing, topic modeling, sentiment analysis, visualization, and interpretation.

The rationale for adopting a computational approach lies in the need to examine large, complex, and fast-evolving social media datasets that cannot be effectively captured through conventional qualitative methods. By employing NLP, this study bridges the methodological gap between political communication, computational linguistics, and digital media studies—providing an inter-disciplinary framework to detect, classify, and interpret manipulative narratives with higher accuracy and replicability. This approach also reflects the study's commitment to research transparency, as NLP allows for systematic documentation of each stage of analysis and ensures reproducibility of findings across platforms.

Data were collected from two platforms. On X, data were obtained through web scraping using IPIFY. Meanwhile, Instagram data were accessed through an ethically approved sampling process, employing purposive and systematic strategies. Ethical considerations were observed by analyzing only publicly available content and ensuring anonymization of user identifiers.

Purposive sampling limited inclusion to comments directly related to political topics and fake accounts, while systematic sampling selected comments at fixed intervals to ensure objectivity and replicability—consistent with guidance on purposive sampling for information-rich cases (Odoh & Olagunju, 2025) and on systematic sampling to preserve representativeness under large populations and restricted access (Campbell et al., 2020).

The dataset included posts, comments, and account metadata. Fake accounts were identified using criteria established by Widyatama and Mahbob (2024). These criteria include: unclear or generic account names that rarely resemble real personal names, often combining random letters and numbers; incomplete profiles with irrelevant or externally sourced profile photos; and content dominated by fabricated narratives suggesting widespread support. Fake accounts typically display abnormal activity patterns, such as mass interactions within short timeframes, repetitive posts with similar topics and styles, minimal original content, frequent reposts, and the circulation of screenshots or hoaxes. Their posts often adopt sensational, provocative, or confrontational tones aimed at provoking emotional responses and polarizing public opinion.

During preprocessing, raw data were cleaned to ensure analytical quality. This stage involved removing stopwords, emojis, URLs, hashtags, and irrelevant characters. Text normalization followed, including lowercasing, tokenization, and issue identification through topic modeling. For this purpose, BERTopic with IndoBERT embeddings was applied, a method proven effective in detecting social media topics (Abdalrazaq et al., 2024). The parameter `nr_topic` was set to 25 to limit topic clusters. BERTopic generated thematic clusters that captured the key issues raised by fake accounts, with a particular focus on clusters related to political issues, which were further examined at the sub-topic level.

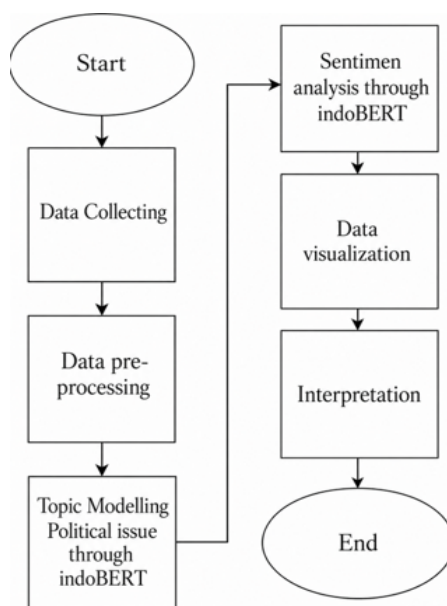
To ensure methodological robustness, sentiment classification was performed using IndoBERT sentiment models fine-tuned for Indonesian text corpora. This allowed precise differentiation of positive, neutral, and negative orientations, reducing linguistic bias and ensuring cultural sensitivity to local political discourse.

Once the political topics had been identified, we performed sentiment analysis to classify account stances toward each issue as positive, negative, or neutral. This enabled a more granular understanding of how fake accounts framed political debate. We then visualized the results using complementary techniques—most notably word clouds to surface dominant keywords—before moving to interpretive analysis. In this final stage, we synthesized the findings by detailing the principal issues targeted by fake accounts, the prevailing sentiment orientations, and the communicative implications for public opinion dynamics and potential polarization.

Methodological validity was strengthened through triangulation across three dimensions: (1) platform comparison (X vs. Instagram), (2) cross-checking narrative clusters with manual inspection of sample posts, and (3) theoretical consistency testing against prior literature on digital astroturfing and emotional diffusion. This triangulation ensures that the computational results align with theoretical expectations and real-world communication behavior.

A schematic diagram of the research workflow and methodological steps is provided in Figure 1. The diagram illustrates the six stages of the study: data collection, data preprocessing, topic modeling, sentiment analysis, visualization, and interpretation. In summary, this methodological framework not only captures the linguistic and emotional dimensions of fake accounts' narratives but also integrates computational analysis with interpretive insights—allowing the study to contribute meaningfully to interdisciplinary debates on political communication, platform governance, and digital education.

Figure 1: Research Workflow and Analytical Design.



3. Findings

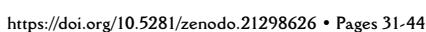
3.1. Fake Accounts' Interaction and Engagement

The dataset comprised 8,102 posts retrieved through NLP (Natural Language Processing), based on topic modeling. Of these, 7,996 were mapped to platform X, while the remaining posts originated from Instagram (IG). Platform identification relied on URLs: X posts consistently included links with the structure x.com/..., whereas IG posts contained instagram.com/... addresses.

Although IG initially appeared to contribute more data, two structural barriers limited the capture of fake account activity. First, IG's API and platform architecture restrict large-scale downloading of comments, making it difficult to comprehensively map fake account behavior in comment sections. Second, most fake IG accounts operate under private settings, limiting visibility of their content, interactions, and narrative distribution. As a result, IG data are comparatively constrained, especially in terms of engagement traces and diffusion networks.

These contrasting results demonstrate that fake accounts are not only disseminators of falsehoods but also strategic users of affordances: X favors wide, reactive, and short-term amplification, while IG enables narrow, sustained, and emotionally resonant diffusion. This evidence underscores that combating disinformation requires differentiated platform governance—targeting algorithmic amplification on open networks and privacy misuse on closed ones.

The research corpus contained 8,102 posts, with 7,996 (98.69%) from X and 106 (1.31%) from IG. This imbalance reflects the structural constraints of IG, where private accounts and limited API access restrict data collection and prevent visibility into many interactions. Conversely, X provides richer interaction traces—replies, retweets, and quotes—that allow more comprehensive mapping of political narratives.



To complement the modeling results, a word cloud visualization illustrates the most frequently used political terms across fake accounts (Figure 2). These clusters represent the core agenda-setting functions of fake accounts. They reveal how manipulative actors reproduce political hierarchies by centering attention on elite figures (“Candidates & Elites”) while simultaneously delegitimizing institutions (“Corruption & Law”). This mirrors global findings where inauthentic networks blend personalized and institutional narratives to maximize emotional salience (Muñoz et al., 2024; Pierri et al., 2020).

The predominance of topics tied to candidates, economic anxiety, and corruption reflects the instrumental use of moral framing—casting political conflict as moral failure or betrayal. Such moral-emotional framing aligns with diffusion theory, where emotionally charged, identity-relevant messages are statistically more shareable (Pröllochs et al., 2021).

3.3. Thematic Landscape of Political Narratives

Topic modeling with BERTopic clustered fake accounts’ narratives into seven main groups (Table 2). From the corpus, two dominant themes emerged. The first, “Candidates & Elites,” comprised 757 posts (9.34% of the corpus; $X = 751$ [99.21%], $IG = 6$ [0.79%]). This theme was characterized by person-centric narratives, ranging from glorification to attacks on political figures such as Jokowi, Prabowo, Gibran, Anies, Ganjar, and Megawati. The second major theme, “Economy & Welfare,” encompassed 482 posts (5.95%; $X = 481$ [99.79%], $IG = 1$ [0.21%]). Here, policy-centric narratives dominated, focusing on inflation, food and fuel prices, subsidies, and poverty—issues frequently invoked to evaluate the government’s performance.

The theme of “Corruption & Law” appeared in 303 posts (3.74%; $X = 301$ [99.34%], $IG = 2$ [0.66%]), often deploying keywords such as KPK (Corruption Eradication Commission), Constitutional Court (MK), laws/regulations (UU/Perppu), and gratification. These narratives typically sought to delegitimize state institutions. “Elections & Campaigns” included 65 posts (0.80%; entirely from X), featuring calls to vote or boycott, as well as references to key electoral moments such as national and regional elections, campaigns, and debates.

The theme of “Identity & Religion” comprised 49 posts (0.60%; $X = 48$ [97.96%], $IG = 1$ [2.04%]), marked by polarization and the use of pejorative labels such as *kadrun* and *cebong* alongside sectarian cues. “Meta-discourse (Hoaxes/Disinformation)” was represented in 46 posts (0.57%; entirely from X), where terms like hoax, slander, and propaganda were used less for fact-checking and more as performative delegitimization of competing narratives. Finally, the “Israel–Palestine” theme appeared in 21 posts (0.26%; all from X), illustrating how global issues were reframed into moral judgments used to critique domestic political positions.

Across themes, three consistent framing strategies emerged: (1). Personalization, which centers on elite figures to anchor public attention; (2). Moralization, which attaches emotional and ethical weight to political disputes; and (3). Delegitimization, which erodes trust in institutions by framing them as corrupt or unjust.

These patterns confirm that fake accounts operate through discursive hybridization—merging emotional, moral, and political vocabularies to sustain polarization. The dominance of “Candidates & Elites” also reflects Indonesia’s highly personalized political culture, where leadership narratives overshadow policy debate (Lim, 2017; Santoso et al., 2020). (Table 2)

Table 2: Dominant Thematic Categories of Fake Accounts Identified by BERTopic.

No	Topic Categories	Lexical choices in posts
1	Elections & Campaigns	Elections, presidential race, and campaigns; focused on mobilizing support.
2	Figures & Elites	Dominant person-centric narratives, such as references to Jokowi, Prabowo, Gibran, Anies, Ganjar, and Megawati
3	Identity Politics & Religion	Dominated by vocabulary such as Islam, caliphate, <i>kadrun</i> , and <i>cebong</i> ; tending to be polarizing.
4	Israel–Palestine Issues	Global narratives are employed as moral frames to interpret and judge domestic politics.
5	Corruption & Law	Characterized by terms such as KPK, Constitutional Court, law/regulations, and corruption; aimed at delegitimizing state institutions.
6	Economy & Welfare	Discussing economy, inflation, food, and subsidies; criticizing government performance.
7	Metadiscourse on disinformation	Characterized by frequent use of terms such as hoax, slander, and disinformation; often employed as labels to delegitimize opponents.

3.5. Sentiment Orientation

The corpus consisted of 8,102 posts, with 7,996 originating from X and 106 from Instagram. Sentiment mapping (Positive–Neutral–Negative) using IndoBERT revealed that on X, 6,801 posts (85.06%) were classified as Neutral, 767 (9.59%) as Positive, and 428 (5.35%) as Negative. On Instagram, the distribution was similar: 99 posts (93.40%) were Neutral, 4 (3.77%) Positive, and 3 (2.83%) Negative. Across both platforms, Neutral sentiment clearly dominated the corpus.

However, deeper analysis revealed that “neutral” posts often contained implicit evaluative cues, irony, or rhetorical framing—a phenomenon common in algorithmic political communication (Törnberg, 2022). Neutrality in form thus masked negativity in tone, which explains the high engagement levels of ostensibly neutral content. When sentiment orientation is examined across the seven thematic clusters, the distribution patterns are summarized in Table 3.

Theme	Thematic Cluster (N)	Positive	Neutral	Negative
Elections & Campaigns	65	14 (21,54%)	42 (64,62%)	9 (13,85%)
Figures & Elites	757	81 (10,70%)	602 (79,52%)	74 (9,78%)
Identity Politics & Religion	49	3 (6,12%)	41 (83,67%)	5 (10,20%)
Israel–Palestine Issues	21	0 (0,00%)	21 (100,00%)	0 (0,00%)
Corruption & Law	303	10 (3,30%)	151 (49,83%)	142 (46,86%)
Economy & Welfare	482	64 (13,28%)	399 (82,78%)	19 (3,94%)
Metadiscourse on disinformation	46	0 (0,00%)	13 (28,26%)	33 (71,74%)

The concentration of high engagement within the “negative” and “morally charged neutral” segments confirms that emotional activation—not factual accuracy—drives diffusion. This is consistent with psychological models of online outrage diffusion (Brady et al., 2021). Moreover, negative sentiment disproportionately targeted institutions, while positive sentiment clustered around populist or identity-linked content, reinforcing ideological divides.

In summary, the findings suggest that fake accounts function as both catalysts and mirrors of political affect. They strategically exploit the affordances of open and closed platforms to trigger selective exposure, emotional contagion, and echo chamber effects. These findings provide empirical grounding for policy recommendations on algorithmic transparency, digital literacy, and media ethics education.

4. Analysis

4.1. Platform Architecture and Narrative Visibility

The data analysis underscores that the quality and heterogeneity of X’s dataset are significantly more complete (including text and interaction metrics such as likes, retweets, replies, and quotes), whereas Instagram (IG) provides only partial data and more limited interactions. This discrepancy highlights a data asymmetry between the two platforms and explains the methodological rationale for conducting more detailed analyses on X. Importantly, these limitations are not attributable to the researchers but are embedded in IG’s platform architecture.

The scholarly implication of this asymmetry is that X offers a more representative lens into open interaction networks (e.g., retweets, quotes, replies), while IG reveals that fake accounts rely on closed networks, using privacy as a “shield” against public detection. This insight contributes beyond data volume—it explains why fake accounts operate differently across platforms.

The corpus reveals a clear dominance of X (7,996 posts; 98.69%) compared with IG (106; 1.31%). This imbalance is not a mere artifact of data collection but rather a structural consequence of platform governance. X leaves rich public interaction traces, whereas IG restricts access to comments and private accounts, leaving many interactions invisible in open data. In this study, the prevalence of private accounts on IG emerged as a major limitation, resulting in thinner engagement footprints compared with X.

Theoretically, these differences align with the framework of digital architectures—the interplay of network structures, interaction features, algorithmic ranking, and datafication models—which shape political communication strategies across platforms (Bossetta, 2018). Prior research shows that politicians and users employ X and IG differently: on X, conversations are elite-driven and easily reach wide audiences, while

IG content is visually curated and emphasizes proximity to followers (Olof Larsson, 2023). From a data governance perspective, official documentation of the Instagram Graph API confirms these restrictions: comment and content access is limited, primarily for professional accounts managed by their owners, and does not extend to private accounts. This explains why IG interactions in this study could not be captured as comprehensively as those on X.

These findings also highlight that the architecture of each platform embodies distinct political and communicative logics. X, as an “open network,” incentivizes performative participation and competitive visibility, which facilitates the viral spread of emotionally charged narratives. IG, conversely, functions as a “closed network” of curated interactions that prioritize aesthetic appeal and personal intimacy. This structural dichotomy directly informs how fake accounts adapt their manipulation strategies: exploiting X’s virality to ignite polarization and using IG’s privacy affordances to maintain echo chambers and long-term ideological cohesion.

From a policy perspective, this architectural asymmetry calls for differentiated regulatory responses. Open-network platforms require algorithmic transparency and virality control (e.g., friction mechanisms to slow the spread of harmful content), while closed-network platforms need stronger detection systems for covert networks that exploit private-group dynamics. Educationally, digital literacy programs must teach users not only to recognize false information but also to understand the architectural incentives that amplify such information.

4.2. Topic Patterns and Framing Strategies

The thematic consistency of our findings resonates with recent advances in short-text topic modeling. For data as concise as tweets or captions—where word co-occurrence is sparse—embedding- and transformer-based methods produce more coherent topics than classical bag-of-words models. Recent scholarship highlights the need for methods that handle contextual scarcity and evaluate topics by substantive usefulness rather than automatic coherence scores (Albalawi et al., 2020).

In the context of mis/disinformation, BERTopic has proven effective for mapping narratives across communities (e.g., COVID-19) and tracking how new issues emerge and diffuse into narrower groups—a parallel to this study’s findings on high-risk themes such as identity and institutional legitimacy (Unlu et al., 2025). For noisy short-text platforms like X/Twitter, staged filtering prior to modeling improves coherence and enhances issue-framing detection, reinforcing the preprocessing steps adopted here (Kant et al., 2025). More recent work combining multimodal neural topic models (e.g., integrating text and images) with misinformation detection further confirms that topic modeling can serve as a central foundation for narrative analysis (Cantini et al., 2025).

Cross-platform framing studies demonstrate that political actors package messages differently depending on the medium, and these strategies affect audience responses: conflict- or morality-framed issues drive retweets, whereas emotionally charged language more often triggers likes (Sahly et al., 2019). This helps explain the dominance of personality-centered and institutional-delegitimization narratives on X compared with IG. Beyond electoral politics, research on social movements identifies three types of framing—diagnostic (problem definition), prognostic (solution/tactics), and motivational (calls to action)—and has built computational models to detect them on Twitter. This aligns with our finding that fake accounts frequently shift from problem labeling (e.g., “corrupt/unconstitutional”) to motivational appeals (e.g., urging voting or boycotts) across different themes (Mendelsohn et al., 2026).

Theoretically, this supports the view that fake accounts operate as “algorithmic activists” who selectively adopt diagnostic and motivational frames to optimize algorithmic rewards. By combining moral-emotional vocabulary with simplified blame attributions, these actors produce discourses that are emotionally engaging but cognitively shallow—precisely the kind of content rewarded by algorithmic curation.

Furthermore, this finding enriches framing theory by introducing a computational dimension: moral-emotional diffusion is not random but strategically synchronized with platform affordances. The blending of identity-based and institutional frames across themes illustrates a hybrid communication logic that merges populist and conspiratorial rhetoric.

This analytical lens allows the study to move beyond content description toward structural interpretation, showing how disinformation systems reproduce power through data architectures, not merely through persuasive language.

4.3. Sentiment Orientation and Diffusion Implications

The sentiment structure of the corpus reveals a dominance of neutral posts across platforms, alongside a small but highly diffusive subset of negative posts tied to specific themes. On X, negative posts ($N=428$) exhibited higher engagement dispersion (median $N=2$; $P75 = 62$) compared with positive ones (median $N=0$; $P75 = 1$). Neutral posts appeared “quiet” at the median ($=1$) but occasionally spiked dramatically (up to 361,900 interactions) when linked to particular actors or events.

This profile aligns with experimental-observational evidence that moral-emotional language and out-group cues are powerful drivers of diffusion in social networks: each additional moral-emotional word increases retweet rates by $\approx 20\text{--}24\%$ on political issues, and out-group language is the strongest predictor of shareability— $\approx 4.8\times$ stronger than negative language and $\approx 6.7\times$ stronger than moral-emotional language itself (Brady et al., 2021).

From a reinforcement-learning perspective, fake accounts appear to “learn” from social feedback: posts that attract likes and retweets are more likely to be repeated and reinforced in subsequent activity (Brady et al., 2021). Even with a small share of negative posts, those carrying emotional or moral weight diffuse more easily and persist longer; an increase of one standard deviation in aggressiveness predicts $\approx 19\%$ more reshares (Pröllochs et al., 2021).

In the Indonesian context, these findings indicate that the “illusion of neutrality” masks deeper ideological polarization. Many posts labeled neutral linguistically still encode implicit antagonism through sarcasm, irony, or moral insinuation—a rhetorical style that has become normalized in Indonesian online discourse (Santoso et al., 2020). This subtle toxicity contributes to what Törnberg (2022) terms *affective polarization*, where emotional hostility persists even in the absence of overt ideological statements.

Literature further shows that false content spreads more widely and quickly than accurate content, especially in political contexts. Crucially, this pattern is not solely attributable to bots—human users are often more eager to share it (Pennycook & Rand, 2019). Within this framework, novelty, moral-emotional charge, and out-group cues operate as the “fuel” for virality on X, while IG’s architecture and governance (privacy and restricted comment access) reduce its susceptibility (Vosoughi et al., 2018).

This suggests that combating online manipulation must integrate technological design, policy regulation, and user education. Platforms should implement sentiment-aware moderation tools that identify emotionally extreme patterns, while governments and educators should emphasize emotional literacy—the ability to recognize manipulative sentiment triggers—as part of digital literacy curricula.

4.4. Interaction Dynamics and the Role of “Super-Spreaders”

Interaction patterns on X are uneven: most posts receive minimal engagement, but a minority explode into viral cascades. This heavy-tailed distribution mirrors global findings on “super-spreaders,” a small group of accounts responsible for disproportionately large shares of interactions. For example, a U.S. study of verified voters found that $\approx 2,107$ accounts ($\approx 0.3\%$ of the sample) generated 80% of shares of unreliable news links during the 2020 election (Allcott et al., 2024). Thus, the spikes in this study’s data are not local anomalies but part of a broader phenomenon.

The identification of super-spreaders in Indonesia’s context reveals how disinformation ecosystems depend on a few high-connectivity nodes. These accounts often masquerade as ordinary citizens yet possess coordinated posting patterns and synchronized language styles—hallmarks of organized influence operations. Detecting these structural “driver nodes” offers policymakers and platforms a leverage point for early intervention.

Such dynamics are explained by cascade models of information diffusion (Sidorov et al., 2025; Zhou et al., 2021), where content spreads widely either through major broadcasts or through multi-generational reshares (Zhang & Ho, 2022). Virality thus depends on initial boosts from accounts with large reach, a finding consistent with prior research showing that cascades emerge when well-networked accounts trigger secondary waves (Goel et al., 2015). Cross-community reshares also amplify reach: retweets that bridge communities are $\approx 1.5\times$ more likely to achieve higher popularity than those confined within a cluster, underscoring the importance of weak ties and network bridges (Tsugawa, 2019).

From a governance standpoint, mapping these cascades has concrete implications: rather than chasing individual posts, regulators and civil society initiatives should focus on disrupting the structural synchrony

among these influential nodes. Network-level intervention—through transparency mandates, coordinated content takedowns, or public exposure of covert amplification networks—can more effectively curb manipulative virality than isolated content removal.

4.5. Cross-National Comparisons and the Indonesian Context

In Indonesia, X clearly dominates political discourse relative to IG, where private accounts and comment restrictions dilute visibility. This finding resonates with comparative studies showing that platform design shapes discourse differently across contexts. For instance, analyses comparing Twitter, Facebook, Reddit, and Gab reveal distinct echo chamber patterns, with Facebook and Twitter fostering more homogeneous exposure compared to Reddit and Gab (Cinelli et al., 2021). In Norway, party leaders use Twitter as a news-driven, public discussion arena, whereas Instagram is leveraged for visual proximity to followers (Skogerbø & Larsson, 2021).

Our findings also align with research in Brazil, where closed channels acted as “warming rooms” for narratives. During the 2018 Brazilian election, disinformation often incubated in private groups before spilling into open platforms such as Twitter (Recuero et al., 2020). This analogy is instructive for Indonesia: when closed channels like IG are difficult to observe, only fragments of orchestration become visible in public data.

In contrast, Indonesia exhibits a distinct hybrid ecology: fake accounts here exploit both open and closed architectures simultaneously—launching viral attacks on X while maintaining continuity and reinforcement through IG. This dual-mode manipulation reflects a higher level of tactical sophistication and suggests that Indonesian digital polarization operates through a multi-platform ecosystem.

This comparative insight contributes to global discussions on political manipulation by demonstrating that platform affordances and local political culture jointly determine disinformation dynamics. For Indonesia, the convergence of algorithmic openness and populist personalization creates a uniquely volatile digital sphere—one that demands integrated responses combining data governance, civic education, and cross-platform collaboration.

5. Conclusion

Based on the analysis, this study concludes that the architecture of social media platforms is a decisive factor shaping the strategies, reach, and impact of information operations. A fundamental distinction exists between X and Instagram (IG): the open, broadcast-centric structure of X facilitates the wide and measurable diffusion of narratives, whereas IG’s closed, visually oriented networks push actors to adopt more covert strategies, using privacy features as shields against public detection. These structural differences directly influence how narratives are framed, with fake accounts strategically emphasizing high-risk topics such as institutional delegitimization and identity politics—issues proven effective in triggering reactions in the public sphere of X.

This study contributes new empirical evidence that the manipulation strategies of fake accounts are not random, but adaptive responses to the affordances of each platform’s architecture. The “open” logic of X rewards visibility and rapid engagement, while the “closed” logic of IG sustains long-term ideological alignment within smaller communities. This duality explains why digital polarization in Indonesia persists even under increasing regulatory and educational interventions.

Another key finding concerns the dynamics of sentiment and engagement. Although the corpus was dominated by neutral content, a disproportionately high diffusion potential was concentrated in a small subset of negative posts. These posts were deliberately framed with moral-emotional language and out-group cues to maximize virality. This mechanism produced a “super-spreader” effect, where a handful of accounts or posts generated massive cascades of interaction, particularly when narratives successfully crossed community boundaries and reached broader audiences. The results demonstrate that disinformation does not spread by volume alone, but by the calculated activation of collective emotions—a pattern consistent with moral-emotional contagion theory and reinforcement-learning behavior in social media systems.

Theoretically, this research advances the concept of *platform architecture as an independent variable* in political communication, showing how technical affordances, governance structures, and cultural contexts jointly determine the reach and tone of manipulative content. It also enriches computational communication studies by integrating NLP-based evidence into interpretive frameworks on digital power and affective polarization.

Practically, the study offers three major implications: (1). For policy makers, the findings highlight the need for algorithmic transparency and virality control mechanisms, especially on open platforms such as X. (2). For platform governance, the results suggest developing detection models that account for architectural differences between open and closed systems, rather than applying one-size-fits-all moderation rules. (3). For education and literacy initiatives, this research underlines the importance of teaching emotional literacy—helping users recognize manipulative language and moral-emotional framing as triggers of polarization.

In Indonesia, where social media constitutes the main arena of political discourse, the findings imply that digital polarization is intensified by the convergence of algorithmic virality and populist communication culture. Addressing this requires collaborative approaches that integrate computational monitoring, civic education, and regulatory innovation.

Limitations of this study include the restricted access to private data on Instagram and the focus on text-based content, which may underrepresent multimodal strategies such as images and videos. Future research should therefore expand to cross-platform and multimodal analyses, combining linguistic, visual, and network-based data to understand the full ecology of disinformation. Longitudinal designs could also reveal how manipulation strategies evolve across electoral cycles and regulatory changes.

In summary, this study emphasizes that strengthening democratic resilience in the digital era requires a dual approach—technological (through transparent algorithms and platform accountability) and educational (through critical and emotional digital literacy). Only by addressing both dimensions can societies effectively mitigate the systemic risks posed by coordinated inauthentic behavior.

Notes

1. The small IG sample reflects platform design: restricted comment access and prevalent private accounts, not researchers' choices.
2. Fake accounts are defined per Widyatama and Mahbob (2024), opaque or generic usernames, thin profiles, repetitive posts, and non-dialogic interactions.

Funding Agency

This research was funded by the Ministry of Higher Education, Science, and Technology – Directorate General of Research and Development, Republic of Indonesia through the Basic Research Scheme, 2025.

References

- Abd-alrazaq, A., Nashwan, A. J., Shah, Z., Abujaber, A., Alhuwail, D., Schneider, J., AlSaad, R., et al. (2024). Machine Learning–Based Approach for Identifying Research Gaps: COVID-19 as a Case Study. *JMIR Formative Research*, 8, e49411. <https://doi.org/10.2196/49411>
- Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, 3, 42. <https://doi.org/10.3389/frai.2020.00042>
- Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., Cisneros, J. C., et al. (2024). The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proceedings of the National Academy of Sciences*, 121(21), e2321584121. <https://doi.org/10.1073/pnas.2321584121>
- Bajari, A., Suryana, A., Wahyudin, U., Mulyana, S., & Rakhmaniar, A. (2023). Hate Speech Through Identity Politics On Social Media Ahead Of The 2024 Presidential Election. *Jurnal Komunikasi*, 15(2), 368–382. <https://doi.org/10.24912/jk.v15i2.25836>
- Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journalism & Mass Communication Quarterly*, 95(2), 471–496. <https://doi.org/10.1177/1077699018763307>
- Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1), 7. <https://doi.org/10.1038/s41467-018-07761-2>
- Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33), eabe5641. <https://doi.org/10.1126/sciadv.abe5641>
- Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., et al. (2020). Purposive sampling: complex or simple? Research case examples. *Journal of Research in Nursing*, 25(8), 652–661. <https://doi.org/10.1177/1744987120927206>
- Cantini, R., Cosentino, C., Kilanioti, I., Marozzo, F., & Talia, D. (2025). Unmasking deception: a topic-oriented multimodal approach to uncover false information on social media. *Machine Learning*, 114(1), 13. <https://doi.org/10.1007/s10994-024-06727-4>
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrocioni, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), e2023301118. <https://doi.org/10.1073/pnas.2023301118>
- Coffé, H. (2017). Citizens' media use and the accuracy of their perceptions of electoral integrity. *International Political Science Review*, 38(3), 281–297. <https://doi.org/10.1177/0192512116640984>

- DeVerna, M. R., Aiyappa, R., Pacheco, D., Bryden, J., & Menczer, F. (2024). Identifying and characterizing superspreaders of low-credibility content on Twitter. *PLOS ONE*, 19(5), e0302201. <https://doi.org/10.1371/journal.pone.0302201>
- Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2015). The Structural Virality of Online Diffusion. *Management Science*, 62(1), 180–196. <https://doi.org/10.1287/mnsc.2015.2158>
- Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5), 472–480. <https://doi.org/10.1038/s41562-020-0833-x>
- Kant, G., Wiebelt, L., Weisser, C., Kis-Katos, K., Lubner, M., & Säfken, B. (2025). An iterative topic model filtering framework for short and noisy user-generated data: analyzing conspiracy theories on twitter. *International Journal of Data Science and Analytics*, 20(2), 269–289. <https://doi.org/10.1007/s41060-022-00321-4>
- Lim, M. (2017). Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia. *Critical Asian Studies*, 49(3), 411–427. <https://doi.org/10.1080/14672715.2017.1341188>
- Mendelsohn, J., Vijan, M., Card, D., & Budak, C. (2026). Framing Social Movements on Social Media: Unpacking Diagnostic, Prognostic, and Motivational Strategies. *Journal of Quantitative Description: Digital Media*, 4, 1–61. <https://doi.org/10.51685/jqd.2024.icwsm.9>
- Mohamad, N. R., Sahab, A., Aribowo, A., Wibisono, R. B., & Asfar, M. (2025). Assessing Political Communication Reform in Digital Age: Case of 2024 Indonesia Presidential Elections. *Journal of Law and Legal Reform*, 6(3), 1631–1684. <https://doi.org/10.15294/jllr.v6i3.20491>
- Muñoz, P., Díez, F., & Bellogín, A. (2024). Modeling disinformation networks on Twitter: structure, behavior, and impact. *Applied Network Science*, 9(1), 4. <https://doi.org/10.1007/s41109-024-00610-w>
- Odoh, A., & Olagunju, A. T. (2025). The Role of Social Media in Combating Corruption, Promoting Accountability, and Sustainable Development in Nigeria's Democracy. *Journal of Developing Societies*, 41(2), 242–267. <https://doi.org/10.1177/0169796x251322248>
- Olof Larsson, A. (2023). The rise of Instagram as a tool for political communication: A longitudinal study of European political parties and their followers. *New Media & Society*, 25(10), 2744–2762. <https://doi.org/10.1177/14614448211034158>
- Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7), 2521–2526. <https://doi.org/10.1073/pnas.1806781116>
- Pierri, F., Artoni, A., & Ceri, S. (2020). Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. *PLOS ONE*, 15(1), e0227821. <https://doi.org/10.1371/journal.pone.0227821>
- Pröllochs, N., Bär, D., & Feuerriegel, S. (2021). Emotions in online rumor diffusion. *EPJ Data Science*, 10(1), 51. <https://doi.org/10.1140/epjds/s13688-021-00307-5>
- Recuero, R., Soares, F., & Vinhas, O. (2020). Discursive strategies for disinformation on WhatsApp and Twitter during the 2018 Brazilian presidential election. *First Monday*, 26(1), 1–16. <https://doi.org/10.5210/fm.v26i1.10551>
- Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolette, L. D. (2023). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health*, 31(7), 1007–1016. <https://doi.org/10.1007/s10389-021-01658-z>
- Rumata, V. M., & Nugraha, F. K. (2020). An Analysis of Fake Narratives on Social Media during 2019 Indonesian Presidential Election. *Jurnal Komunikasi: Malaysian Journal of Communication*, 36(4), 351–368. <https://doi.org/10.17576/JKMJC-2020-3604-22>
- Sahly, A., Shao, C., & Kwon, K. H. (2019). Social Media for Political Campaigns: An Examination of Trump's and Clinton's Frame Building and Its Effect on Audience Engagement. *Social Media + Society*, 5(2), 2056305119855141. <https://doi.org/10.1177/2056305119855141>
- Santoso, D. H., Aziz, J., Utari, P., & Kartono, D. T. (2020). Populism in New Media: The Online Presidential Campaign Discourse in Indonesia. *Gema Online Journal of Language Studies*, 20(2), 115–133. <https://doi.org/10.17576/gema-2020-2002-07>
- Sharma, R. K., Bharathy, G., Karimi, F., Mishra, A. V., & Prasad, M. (2023). Thematic Analysis of Big Data in Financial Institutions Using NLP Techniques with a Cloud Computing Perspective: A Systematic Literature Review. *Information*, 14(10), 577. <https://doi.org/10.3390/info14100577>
- Sidorov, M., Hadar, O., & Vilenchik, D. (2025). Revisiting Information Cascades in Online Social Networks. *Mathematics*, 13(1), 77. <https://doi.org/10.3390/math13010077>
- Skogerbø, E., & Larsson, A. O. (2021). Comparing Twitter and Instagram as platforms for party leader communication: Findings from the 2017 Norwegian election. *First Monday*, 26(10), 1–18. <https://doi.org/10.5210/fm.v26i10.10638>
- Subekti, D., Yusuf, M., Saadah, M., & Wahid, M. (2025). Social media and disinformation for candidates: the evidence in the 2024 Indonesian presidential election. *Frontiers in Political Science*, 7, 1625535. <https://doi.org/10.3389/fpos.2025.1625535>
- Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42), e2207159119. <https://doi.org/10.1073/pnas.2207159119>
- Tsugawa, S. (2019). Empirical Analysis of the Relation between Community Structure and Cascading Retweet Diffusion. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(1), 493–504. <https://doi.org/10.1609/icwsm.v13i01.3247>
- Unlu, A., Truong, S., Sawhney, N., & Tammi, T. (2025). Setting the misinformation agenda: Modeling COVID-19 narratives in Twitter communities. *New Media & Society*, 27(7), 3973–3997. <https://doi.org/10.1177/14614448241232079>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wendratama, E., Aprilia, M. P., Sari, Y. A., Kurnia, N., & Utomo, W. P. (2023). Domesticating WhatsApp Groups: Indonesian Women's Experience with Misinformation and Hate Speech in the 2019 Election. *Jurnal Komunikasi: Malaysian Journal of Communication*, 39(4), 372–390. <https://doi.org/10.17576/JKMJC-2023-3904-20>
- Widyatama, R., & Mahbob, M. H. (2024). The Potential Hazards of Fake Accounts and Buzzer Behaviour on Deliberative Democracy. *Jurnal Komunikasi: Malaysian Journal of Communication*, 40(1), 324–341. <https://doi.org/10.17576/JKMJC-2024-4001-18>

- Yerlikaya, T., & Aslan, S. T. (2020). Social Media and Fake News in the Post-Truth Era: The Manipulation of Politics in the Election Process. *Insight Turkey*, 22(2), 177–196. <https://doi.org/10.25253/99.2020222.11>
- Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society*, 23(5), 1080–1098. <https://doi.org/10.1177/1461444820908530>
- Zhang, X., & Ho, J. C. F. (2022). Exploring the Fragmentation of the Representation of Data-Driven Journalism in the Twittersphere: A Network Analytics Approach. *Social Science Computer Review*, 40(1), 42–60. <https://doi.org/10.1177/0894439320905522>
- Zhou, F., Xu, X., Trajcevski, G., & Zhang, K. (2021). A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Computing Surveys (CSUR)*, 54(2), 1–36. <https://doi.org/10.1145/3433000>