Indonesian Journal of Language and Communication Studies
Vol. 1, No. 6, June 2026, pp. 104~115
ISSN: 3163-6209, DOI:

From moderation to marginalization: language, power, and content governance on Indonesian social media platforms

Reza Pahlawan
Universitas Negeri Yogyakarta, Indonesia

Article Info

Article history:

Received Apr 25, 2026
Revised May 22, 2026
Accepted Jun 30, 2026

Keywords:

content governance;
linguistic marginalization;
platform power;
social media;
transparency.

ABSTRACT

Background: Content moderation increasingly governs linguistic visibility, public participation, and institutional authority across Indonesia's multilingual social-media environment, where standardized harm categories may inadequately recognize contextual, regional, and identity-based expression. **Objective:** This study examines how regulatory language, platform policies, enforcement procedures, and civil-society critiques configure relationships among language, power, and marginalization in Indonesian content governance. **Method:** Using a qualitative corpus-assisted design, this study analyses 26 verified public documents from 2020–2026, integrating official regulations, platform standards, enforcement communications, civil-society assessments, and comparative governance materials across Indonesian public digital communication ecosystems, through critical policy discourse analysis, moderation-process mapping, and multimodal governance analysis. **Results:** Findings show that harm definition and contextual classification dominate governance discourse, while linguistic accessibility and culturally situated interpretation receive comparatively limited institutional attention. Enforcement documents prioritize automated detection, takedown, and compliance, whereas explanations, appeal mechanisms, and unequal consequences remain less systematically articulated. **Implication:** Civil-society and international materials consequently reposition transparency, proportionality, participation, localization, and independent review as necessary correctives to concentrated state-platform authority. **Novelty:** This study contributes an integrated account of moderation as a distributive governance system, demonstrating that marginalization may emerge not solely from content removal but from institutional asymmetries between classificatory power, visibility control, contextual competence, and users' capacity to understand and contest decisions.

Corresponding Author:

Reza Pahlawan
Universitas Negeri Yogyakarta, Indonesia
Jl. Colombo No.1, Karangmalang, Caturtunggal, Kec. Depok, Sleman, Yogyakarta, 55281, Indonesia
Email: rezapahlawan@uny.ac.id

1. INTRODUCTION

Indonesia's social-media environment has become a site where linguistic expression, public safety, and institutional authority intersect. In 2024, Indonesia counted 221,563,479 internet users, equivalent to 79.5% of its population, giving platform rules extensive reach across political, cultural, and everyday communication. Within this environment, harmful speech is neither marginal nor linguistically simple: Indonesian online discourse combines standard Indonesian, regional languages, slang, code-switching, irony, and culturally situated insults that complicate automated classification [1], [2]. Moderation therefore extends beyond removing demonstrably illegal material; it determines whose speech remains searchable,

recommendable, monetizable, or appealable. Indonesian authorities have expanded compliance mechanisms through platform registration, takedown obligations, and accelerated responses to designated urgent content. Such arrangements make content governance a distributive institution: decisions framed as protection can allocate visibility, credibility, and procedural access unevenly. Examining this infrastructure is necessary because opaque enforcement may transform moderation from harm reduction into marginalization.

Existing scholarship has established components of this problem without integrating them into an Indonesia-centred account of language and governance. Computational research maps progress and limitations in Indonesian hate-speech and abusive-language detection, particularly contextual ambiguity, dataset quality, linguistic variation, and category definition [1], [3], [4]. Governance research examines how guidelines become enforcement decisions, how platforms communicate legitimacy, and how human-rights principles constrain private rulemaking [5], [6], [7]. Studies of marginalized users document disproportionate removals, invisible sanctions, uncertain appeals, and algorithmic suppression affecting creators, transgender users, disabled users, and vulnerable communities [8], [9], [10]. Indonesian studies, however, predominantly address toxicity detection, hate-speech forms, religious moderation, or regulatory effectiveness separately. They rarely connect linguistic normativity, state-platform power, visibility outcomes, and counter-governance within one verifiable corpus of policies, enforcement statements, platform procedures, and civil-society critiques.

This study investigates how moderation becomes a mechanism through which language is classified, authority is exercised, and unequal visibility is potentially produced on social-media platforms operating in Indonesia. It asks three questions. First, how do Indonesian regulations and platform policies define harmful, acceptable, exceptional, and context-dependent expression, and what assumptions about language underpin those definitions? Second, how do governance procedures—detection, reporting, notification, removal, restriction, demonetization, appeal, and disclosure—distribute power among government agencies, platform companies, users, and affected communities? Third, how do civil-society and rights-based documents challenge dominant moderation models and formulate alternatives grounded in transparency, participation, localization, and procedural fairness? To answer these questions, this study analyses a verifiable corpus comprising regulations, governmental enforcement communications, platform standards and appeal procedures, civil-society assessments, and governance materials published or active between 2020 and 2026. This design treats policy language, enforcement architecture, and public contestation as distinct but connected layers of content governance.

This study advances the provisional argument that marginalization emerges from the interaction among linguistic standardization, enforcement asymmetry, and limited procedural intelligibility. Global moderation categories may appear neutral yet remain insufficiently responsive to Indonesian pragmatic meanings, regional-language resources, reclaimed expressions, satire, defensive speech, and politically situated context. Accelerated takedown systems and opaque visibility controls can privilege institutional definitions of urgency over users' ability to understand or contest decisions. Research suggests governance is experienced unevenly when moderation combines automated classification, informal suppression, and restricted remedies [9], [11], [12]. This study tests whether Indonesian content governance constructs a hierarchy of recognizable speech: language readily legible to regulators and platforms receives procedural treatment, whereas context-dependent or marginalized expression bears greater interpretive risk. Its broader implication is that effective moderation requires governance redesigned around linguistic competence, explainable enforcement, accessible appeal, and accountable participation in contemporary Indonesia.

2. LITERATURE REVIEW

2.1. Content moderation

Content moderation refers to institutional practices through which platforms classify, restrict, remove, or otherwise govern user-generated expression. Technical accounts emphasize detection accuracy, policy enforcement, and scalability, whereas governance scholarship treats moderation as rulemaking that structures participation, rights, and legitimacy [3], [5], [13]. Human-rights approaches further reject a purely operational definition, arguing that moderation redistributes communicative opportunities and may reproduce ideological inequality [6], [14]. Thus, moderation should be understood not as neutral filtering but as a socio-technical institution connecting linguistic classification, corporate authority, state regulation, and user vulnerability.

Its principal dimensions include rule formulation, detection, enforcement, visibility management, notification, appeal, and transparency. Detection may involve automated classifiers or human review, while enforcement ranges from removal and account suspension to demonetization, warning labels, and algorithmic downranking [15], [16]. Research also distinguishes explicit sanctions from invisible moderation, where users experience reduced reach without intelligible explanation [9], [12]. Additional indicators include contextual sensitivity, linguistic competence, proportionality, and remedy. These dimensions demonstrate

that moderation quality depends not only on identifying harmful content but also on explaining decisions and enabling contestation.

2.2. Platform power

Platform power denotes the capacity of digital intermediaries to define legitimate expression, allocate visibility, and shape public participation through infrastructures they privately control. Political-economy perspectives foreground ownership, advertising incentives, and market concentration, whereas governance perspectives emphasize institutional rulemaking and delegated regulatory authority [17], [18]. Discursive approaches add that platforms secure legitimacy by presenting enforcement as balanced, neutral, and safety-oriented, even when users encounter contradictory outcomes [7], [19]. Platform power is therefore neither exclusively economic nor algorithmic; it operates through policies, interfaces, classifications, communications, and selective disclosure that normalize corporate authority over speech.

Indicators of platform power include agenda setting, category definition, data control, enforcement discretion, algorithmic ranking, and control over procedural knowledge. Platforms determine what counts as hate, harassment, misinformation, or acceptable contextual speech, yet users often cannot inspect classifiers, evidence, or distributional interventions [20], [21]. Power also appears in asymmetrical appeals, standardized interfaces, and corporate communications that shift responsibility onto users while obscuring institutional error [19], [22]. Consequently, analysis must trace how authority moves from policy language to technical enforcement and how opacity converts platform discretion into durable governance capacity today.

2.3. Marginalization in content governance

Marginalization in content governance describes patterned disadvantage produced when moderation systems misrecognize, disproportionately sanction, or render certain users and expressions less visible. Some studies define it through unequal removal rates, while others emphasize cumulative harms involving identity suppression, inaccessible appeals, and uncertainty about platform norms [8], [11]. Research on blind TikTok users shows that defensive or contextual speech may be penalized when systems fail to interpret disability-related harassment [10]. Linguistic studies indicate that abusive-language detection remains vulnerable to contextual ambiguity in Indonesian discourse [1], [2].

Its indicators include disproportionate removal, reduced visibility, identity erasure, contextual misclassification, procedural burden, and exclusion from governance design. Marginalization can also emerge when local languages, reclaimed terms, satire, or oppositional discourse are less legible to standardized moderation systems [23], [24]. Yet exclusion is not inevitable: user-centered policy resources, personalized moderation choices, independent review, and reparative approaches may redistribute interpretive authority [11], [15], [25]. This study therefore positions marginalization as the relational outcome of linguistic normativity, concentrated platform power, and unequal procedural access within Indonesian content governance in practice today.

3. METHOD

This study examined 26 publicly accessible documents as its material object and treated each document, policy clause, enforcement statement, procedural explanation, and short visual unit as an analytical unit. Corpus composition comprised 15 primary documents, six secondary documents, and five contextual documents published or active between 24 November 2020 and 14 June 2026. Primary materials included Indonesian regulations, government enforcement communications, Meta policies, and YouTube procedures. Secondary materials consisted of SAFEnet and ELSAM analyses, while contextual materials included UNESCO frameworks and public contestation documents. Together, these sources captured regulatory language, platform rules, enforcement mechanisms, rights claims, and marginalization discourses comparatively.

This study adopted a qualitative, corpus-assisted policy discourse design combining critical policy discourse analysis, process tracing, and multimodal governance analysis. This design was systematically selected because content moderation operates through language, institutional procedure, technical intervention, and public justification. Rather than measuring causal effects, this study compared how actors defined harm, allocated authority, specified sanctions, and constructed procedural legitimacy across state, platform, and civil-society texts. Cross-document comparison enabled identification of convergences and tensions among Indonesian regulation, global platform standards, and rights-based critiques. The design therefore linked normative classification with enforcement architecture without treating policy statements as evidence of users' lived outcomes.

Information sources were limited to identifiable and publicly verifiable publishers. Government materials came from Komdigi and its legal documentation portal; platform materials came from Meta Transparency Center and YouTube Help; civil-society materials came from SAFEnet and ELSAM; and

comparative context came from UNESCO. Scientific interpretation was supported by recent scholarship on moderation governance, platform power, linguistic classification, and marginalized users, including Singhal et al. (2022), Griffin (2023), Ibrohim and Budi (2023), Mayworm et al. (2023), and Lyu et al. (2024) [1], [5], [6], [10], [11]. Accordingly, absolutely no anonymous repost, private account, unsupported screenshot, inaccessible page, or unverifiable claim entered this carefully delimited research corpus.

Table 1. Dataset Composition

Code	Source	Scope/location	Period	Number	Main characteristics
PRI-GOV	Komdigi and JDIH Komdigi	Indonesia, national	2020–2026	8	Regulations, enforcement announcements, compliance statements, sanctions, and SAMAN implementation
PRI-META	Meta Transparency Center	Global policies applicable in Indonesia	Current policies retrieved in 2026	5	Community Standards covering hateful conduct, harassment, misinformation, and violent content
PRI-YT	YouTube Help	Global policies applicable in Indonesia	Current procedures retrieved in 2026	2	Content restrictions, contextual exceptions, suspension, monetization, and appeal procedures
SEC-SAFE	SAFEnet	Indonesia, national	2023–2024	3	Human-rights moderation, vulnerable groups, online attacks, and platform accountability
SEC-ELSAM	ELSAM and civil-society coalition	Indonesia, national	2020–2026	3	Legal critique, freedom of expression, proportionality, and platform governance
CTX-UNESCO	UNESCO	Indonesia and comparative settings	2023–2024/current	3	International governance framework, harmful content, local languages, and multistakeholder responses
CTX-SAFE	SAFEnet	Indonesia, national	2025	2	Public contestation, censorship concerns, transparency, and resistance to SAMAN
Total			2020–2026	26	Primary, secondary, and contextual documentary corpus

Data collection proceeded through structured web searching using Indonesian and English terms related to content moderation, platform governance, harmful speech, takedown, appeal, transparency, vulnerable groups, local languages, and SAMAN. Each eligible item was recorded with a unique identifier, publisher, title, date, source type, platform scope, region, language, URL, retrieval date, verification status, quotation context, and relevant excerpt. Visual or audiovisual items were assigned fields for image description, actor, symbol, colour, layout, logo position, frame, timestamp, transcript, and screenshot reference. Duplicate records were removed through exact-URL matching, normalized-title comparison, and institution-date-title checking before coding and interpretive comparison across three analytical instruments.

Analysis followed three procedures. First, critical policy discourse analysis coded regulatory obligation, linguistic accessibility, harm definition, legitimation, contextual exception, and discursive silence. Second, a moderation process and power matrix traced rule formulation, detection, notification, sanction, visibility outcome, appeal, remedy, and transparency. Third, multimodal governance analysis examined how authority, vulnerability, urgency, and compliance were represented visually and verbally. Initial deductive codes were refined through iterative comparison across primary, secondary, and contextual materials. Interpretive claims were retained only when supported by textual or visual evidence, while disagreements among sources were treated as analytically significant rather than resolved through averaging or causal inference.

4. RESULTS

4.1. Linguistic normativity and the architecture of content governance

Language entered content governance through five interrelated operations: obligation, accessibility, classification, legitimation, and contextual differentiation. As shown in Figure 1, across 26 documents, harm definition appeared in eight items or 30.8%, followed by contextual exception in seven items or 26.9%. Regulatory obligation occurred in five documents, while linguistic accessibility and institutional legitimation each appeared in four. These distributions indicate that governance documents devote greater explicit attention to naming prohibited or harmful expression than to explaining linguistic diversity or institutional justification. Primary sources dominated obligation and classification codes, whereas contextual documents contributed more strongly to exceptions and localization concerns. Such variation supports treating content moderation as a layered policy system rather than a unified set of technical rules, consistent with governance accounts emphasizing movement from guidelines to institutional enforcement [5], [13].

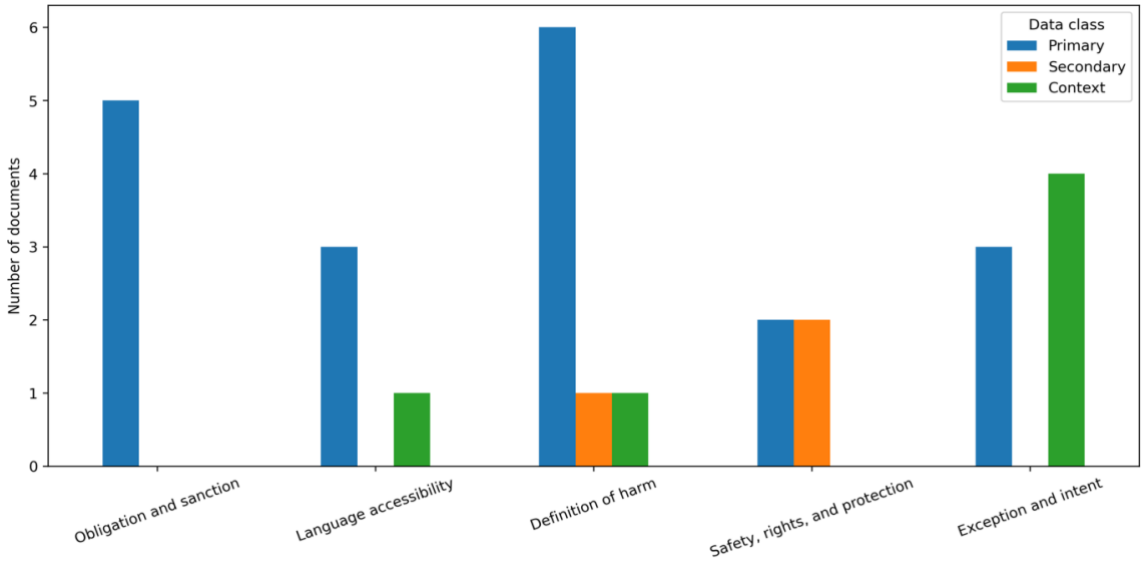

Figure 1. Distribution of linguistic-normativity codes by data class

Content classification formed the most visible pattern, particularly within government regulations and platform policies. In Figure 1, six of eight documents coded for harm definition were primary materials, demonstrating that official actors retain substantial authority to define hate, harassment, violence, misinformation, obscenity, and prohibited information. Contextual differentiation was distributed differently: four of seven occurrences appeared in contextual materials, while three appeared in platform policies. Regulatory documents rarely elaborated educational, documentary, artistic, defensive, or public-interest exceptions with comparable specificity. Linguistic accessibility was also limited, occurring in only four documents despite Indonesia's multilingual environment. One regulation framed Indonesian as a service-access requirement, while later materials identified local-language capacity as an unresolved institutional issue. Objectively, this pattern shows an asymmetry between detailed prohibition and less developed treatment of language-specific context, interpretive variation, and exceptions.

This asymmetry suggests that linguistic normativity is produced through differential institutional recognition rather than through explicit language prohibition alone. Platforms and regulators make expression governable by translating contextual communication into standardized categories of harm, urgency, target, and sanction. Such categorization is necessary for moderation at scale, but its apparent neutrality can conceal interpretive assumptions about directness, intention, identity, satire, and culturally situated meaning [1], [20]. Human-rights scholarship similarly argues that content categories become distributive mechanisms when institutional actors control definitions while affected users possess limited interpretive or procedural authority [6], [14]. This study therefore understands linguistic marginalization as a risk arising when classificatory precision for prohibited speech exceeds institutional capacity to recognize multilingual context, legitimate exceptions, and community-specific meanings.

4.2. Enforcement opacity, unequal visibility, and marginalization

Moderation procedures were distributed unevenly across detection, enforcement, visibility management, explanation, remedy, and social impact. Table 2 shows that automated detection appeared in

ten documents, representing 38.5% of the corpus, while removal and takedown occurred in seven documents or 26.9%. Unequal impact was identified in four documents, whereas visibility reduction appeared in three. Procedural opacity and appeal mechanisms each occurred in only two documents. Primary sources concentrated on detection, sanctions, and platform obligations, while secondary and contextual materials documented opacity and unequal consequences. This distribution reveals a procedural imbalance: institutional sources explain how content should be identified and restricted more frequently than how affected users can obtain explanations, challenge classifications, or demonstrate context. Moderation is therefore documented more fully as an enforcement sequence than as a reciprocal system of accountability.

Table 2 shows that detection and removal formed the dominant operational pattern, but they were not accompanied by equally detailed accounts of remedy or disclosure. Seven of ten automated-detection references appeared in primary documents, indicating strong institutional emphasis on technological capacity, rapid reporting, and compliance. Removal was similarly concentrated in regulations and government enforcement communications. By contrast, all three visibility-management references came from platform policies, where warning labels, age restrictions, and distribution controls were distinguished from deletion. Unequal impact was recorded exclusively in secondary and contextual sources, particularly documents addressing vulnerable users, local-language communities, and rights-based governance. Procedural opacity followed the same pattern, appearing only through external assessment. Accordingly, official materials foreground intervention, while civil-society and international materials supply most descriptions of uncertainty, exclusion, and distributive harm.

This divergence supports an understanding of platform power as control over both communicative visibility and procedural knowledge. Removal is readily observable, but downranking, demonetization, automated flagging, and recommendation exclusion may remain difficult for users to identify or contest. Duffy and Meisner (2022) describe such conditions as algorithmic invisibility [9], while Thach et al. (2022) show how marginalized users develop informal explanations for moderation when official information is unavailable [12]. Haimson et al. (2021) similarly demonstrate that enforcement burdens are not evenly distributed across user groups [8]. Within this study, marginalization is therefore not inferred from every restrictive action. It arises when detection and sanction are institutionally detailed, yet explanation, contextual review, and remedy remain comparatively underdeveloped. This asymmetry allows moderation to operate as governance without providing affected users equivalent interpretive authority.

Table 2. Enforcement opacity, visibility management, and unequal moderation outcomes

Main category	Subcategory	Operational indicator	Frequency (n = 26)	Percentage	Data variation: primary/secondary /context	Evidences	Data codes
Detection infrastructure	Automated detection	AI-supported identification, automated flagging, monitoring systems, or machine-assisted classification	10	38.5%	7/2/1	Stronger content-moderation systems and rapid reporting procedures were presented as responses to AI-generated sexual content.	PRI-GOV-ID-2025-003–006; PRI-GOV-ID-2026-007–008; PRI-PLT-META-000-012; SEC-CSO-SAFE-2024-017; SEC-CSO-ELSAM-2026-019; CTX-IGO-UNESCO-2023-022
Enforcement action	Removal and takedown	Deletion, blocking, access termination, geoblocking, or compelled platform removal	7	26.9%	4/2/1	Platforms are required to respond to governmental notifications concerning prohibited or urgent content within prescribed periods.	PRI-GOV-ID-2020-001; PRI-GOV-ID-2025-003; PRI-GOV-ID-2025-005–006; SEC-CSO-ELSAM-2020-021; SEC-CSO-ELSAM-2026-020; CTX-CSO-SAFE-2025-026
Visibility management	Restriction and downranking	Warning labels, age restrictions, demonetization, reduced distribution, or continued availability under limitation	3	11.5%	3/0/0	Meta removes highly graphic content while applying warning labels to material considered less severe.	PRI-PLT-META-000-012–013; PRI-PLT-YT-000-014
Procedural governance	Opacity and insufficient explanation	Unclear standards, unavailable evidence, limited disclosure, or uncertainty concerning institutional responsibility	2	7.7%	0/1/1	Limited information regarding harmful content and governance effectiveness restricts evidence-based assessment.	SEC-CSO-ELSAM-2026-019; CTX-IGO-UNESCO-2023-022
Procedural remedy	Appeal, review, and correction	Formal appeal, reapplication, review, designated contact point, or decision correction	2	7.7%	2/0/0	YouTube provides appeal and reapplication routes for selected monetization suspensions and rejected applications.	PRI-GOV-ID-2025-006; PRI-PLT-YT-000-015
Distributional consequence	Unequal impact	Disproportionate burden, misclassification, identity suppression, or harm affecting vulnerable and marginalized groups	4	15.4%	0/2/2	Moderation failures may expose vulnerable groups to discrimination, stigma, and recurrent disinformation.	SEC-CSO-SAFE-2023-016; SEC-CSO-SAFE-2024-018; CTX-IGO-UNESCO-2023-022; CTX-IGO-UNESCO-2024-024

4.3. Counter-Governance, Accountability Claims, and Rights-Based Alternatives

Challenges to established moderation arrangements were organized around transparency, human rights, participation, institutional collaboration, resistance, and localization. Figure 2 figures out that human-rights and proportionality critiques appeared most frequently, occurring in five documents or 19.2% of this corpus. Transparency and multistakeholder governance each appeared in three documents, while digital resistance occurred in two. Participation deficit and localization demand were each recorded once. Secondary and contextual sources dominated these categories, particularly rights-based criticism and collaborative alternatives, whereas primary sources contributed mainly to transparency procedures and localized enforcement capacity. This distribution indicates that official documents primarily articulate compliance and intervention, while external actors provide most normative evaluations of whether moderation is proportionate, participatory, explainable, and attentive to local communicative conditions.

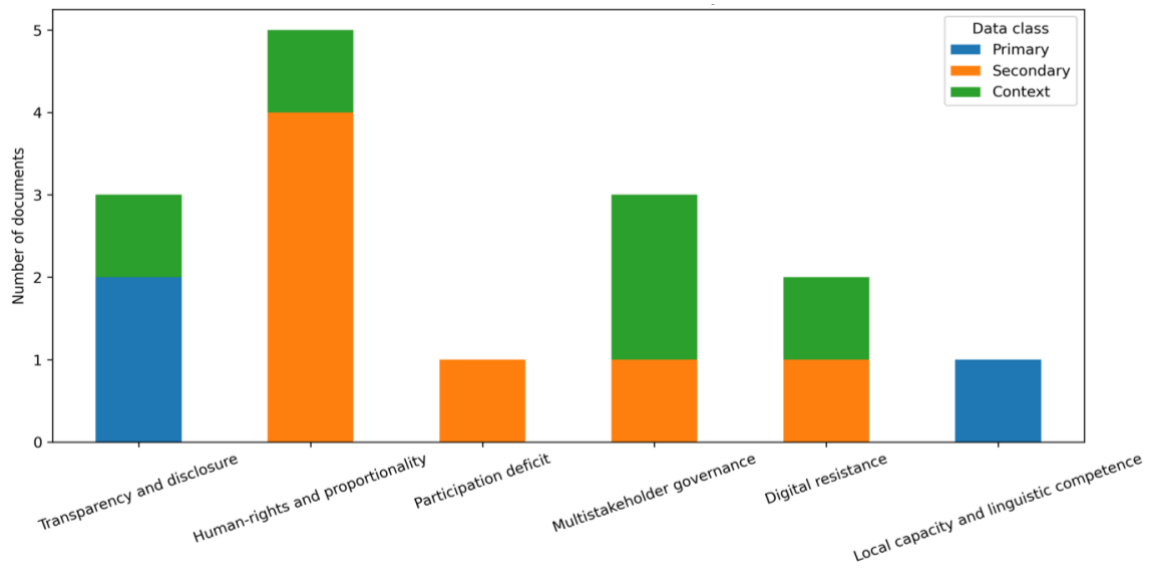

Figure 2. Distribution of counter-governance codes by data class

As shown in Figure 2, human-rights critique constituted the clearest counter-governance pattern, with four of five occurrences located in civil-society documents. These texts connected freedom of expression with due process, proportionality, protection from harm, and legal accountability rather than treating rights as an argument against moderation itself. Multistakeholder alternatives appeared mainly in UNESCO and SAFEnet materials, where collaboration, research, literacy, and affected-community participation were presented as institutional correctives. Transparency claims focused on disclosure of requests, legal grounds, platform responses, and enforcement outcomes. Digital resistance was less frequent but more explicitly oppositional, framing expansive or removal-centred governance as a censorship risk. Participation and localization appeared only sparsely, suggesting that these principles remain less operationalized than broad rights language within available public documentation.

These patterns reposition counter-governance as a struggle over who may define harm, review evidence, and authorize communicative restriction. Griffin (2023) and Sander (2021) argue that rights in platform governance cannot be secured through corporate discretion alone [6], [18], while Heldt and Dreyer (2021) identify independent decision-making bodies as possible institutional safeguards [22]. Peterson-Salahuddin (2024) similarly proposes reparative approaches capable of addressing accumulated moderation harms rather than merely improving classifier accuracy [25]. Within this study, rights-based alternatives are therefore understood as redistributions of interpretive and procedural authority. Transparency makes enforcement visible, participation broadens rulemaking, multistakeholder review disperses institutional control, and localization strengthens linguistic competence. Counter-governance consequently does not imply unregulated speech; it describes moderation reorganized around accountability, proportionality, remedy, and meaningful representation.

5. DISCUSSION

The first implication concerns moderation's promise and distributive limits. By concentrating on harm definition more than on linguistic accessibility or contextual differentiation, governance systems become efficient at classifying speech but less capable of recognizing why expressions mean differently across communities. This asymmetry matters because moderation is not only a protective filter; it is also an infrastructure that allocates legitimacy, reach, and recognition. Findings align with Ibrohim and Budi (2023), who identify contextual ambiguity as a persistent challenge in Indonesian hate-speech detection [1], and with Oh and Downey (2024), who question whether toxic-language systems can sustain democratic discourse [20]. Yet this pattern does not imply that classification itself is inherently exclusionary. Gongane et al. (2022) show that scalable detection remains necessary for addressing harmful content [3]. Dysfunction emerges when standardized categories operate without equivalent mechanisms for interpretation. Consequently, moderation can protect users while narrowing the range of speech institutions can competently understand.

This asymmetry is structurally produced by the need to convert heterogeneous communication into administratively manageable categories. Regulators and platforms require definitions that can guide enforcement across vast volumes of content, but scalability rewards lexical clarity, standardized thresholds, and repeatable procedures rather than pragmatic nuance. Such conditions create a correlation between institutional efficiency and interpretive reduction, particularly in multilingual environments. Singhal et al. (2022) describe this gap between guidelines and enforcement [5], while Endres et al. (2023) argue that content-management bias becomes consequential when formal categories obscure unequal effects [14]. Griffin (2023) similarly links rights allocation to ideological assumptions embedded in governance design [6]. In this study, the underlying structure is not simply technological inadequacy; it is the institutional prioritization of actionable certainty over contextual complexity. Where local language, satire, reclaimed speech, or defensive expression demands interpretive labor, systems may default to standardized readings. Marginalization thus emerges through administrative simplification rather than explicit exclusion.

The second implication concerns the difference between visible enforcement and invisible governance. Removal and automated detection were institutionally prominent, whereas appeals, explanations, and unequal impacts appeared less frequently. This pattern indicates that moderation functions as a mechanism of intervention but less consistently as a system of reciprocal accountability. The consequence is that users may experience restrictions without access to knowledge required to assess, contest, or correct them. This interpretation supports Duffy and Meisner's (2022) account of algorithmic invisibility [9] and Thach et al.'s (2022) finding that marginalized users develop informal theories when platform decisions remain opaque [12]. It also resonates with Divon et al. (2025), who conceptualize corporate communication as platform gaslighting when official explanations contradict user experience [19]. However, Jhaver et al. (2023) demonstrate that users may value tailored moderation choices, suggesting that governance need not remain uniformly opaque [15]. Its dysfunction lies in unevenly institutionalized explanation and remedy, rather than moderation itself.

The underlying structure of this opacity is a concentration of informational and procedural resources within platforms and regulatory institutions. Platforms control policy interpretation, classifier design, evidence logs, visibility systems, and appeal interfaces, while users generally encounter only outcomes. This asymmetry converts technical discretion into governance power because those who impose classifications also determine what can be known about them. Rogers (2021) shows how platform architectures privilege certain information [21], while Liu (2025) explains how legitimacy is communicated through claims of balance and neutrality [7]. Haimson et al. (2021) further demonstrate that moderation gray areas are experienced differently across social groups [8]. In this study, unequal visibility is correlated with unequal access to procedural knowledge. Downranking, demonetization, and automated flagging are especially consequential because they may not generate clear evidence of intervention. The deeper problem is a governance structure in which institutional actors retain interpretive authority while affected users bear uncertainty and proof burdens.

The third implication concerns the normative role of counter-governance. Rights-based critiques, transparency demands, and multistakeholder alternatives indicate that opposition to current moderation arrangements should not be read as opposition to content regulation itself. Instead, these interventions seek to redesign who participates, what must be disclosed, and how restrictive decisions are reviewed. This distinction is important because debates often polarize around safety versus expression, whereas the findings reveal demands for both protection from harm and stronger procedural safeguards. Sander (2021) and Griffin (2023) similarly argue that human rights cannot be reduced to either unrestricted speech or corporate compliance [6], [18]. Heldt and Dreyer (2021) propose independent bodies as potential correctives [22], while Peterson-Salahuddin (2024) advances reparative approaches to accumulated moderation harms [25]. Counter-governance therefore functions as an institutional reform agenda. Its significance lies in shifting

moderation from unilateral control toward accountable, reviewable, and locally informed decision-making without abandoning legitimate harm prevention or broader public-safety obligations.

This reform agenda emerges because state-platform governance is structurally bilateral, whereas moderation consequences are socially distributed. Governments define legal obligations, platforms operationalize enforcement, but users, marginalized communities, and civil-society actors absorb communicative risks without equivalent authority over rules or remedies. Mayworm et al. (2023) show that marginalized users construct platform "folk theories" when official governance excludes their knowledge [11], while Mayworm et al. (2024) demonstrate the value of resources designed around affected communities [26]. Lyu et al. (2024) likewise reveal how disability-related context can be misread when moderation lacks situated understanding [10]. In this study, transparency, participation, localization, and independent review are interconnected because each redistributes interpretive power. Transparency exposes decisions, participation broadens rule formation, localization improves contextual competence, and review constrains institutional discretion. Their limited presence in official documents suggests that counter-governance remains more aspirational than embedded. Even so, it offers a viable route from moderation as control toward accountable public-interest governance.

6. CONCLUSION

This study demonstrates that content moderation in Indonesia functions not merely as technical filtering but as a layered regime of linguistic classification, visibility allocation, and institutional authority. Its lesson is that marginalization emerges when harm categories, enforcement speed, and automated detection develop more rapidly than contextual interpretation, procedural explanation, and accessible remedy. Methodologically, this study contributes an integrated framework combining critical policy discourse analysis, moderation-process mapping, and multimodal governance analysis across regulatory, platform, civil-society, and contextual documents. Conceptually, it reframes moderation as a distributive system in which language, power, and accountability jointly shape who remains visible, intelligible, and contestable online.

Several limitations qualify these conclusions. This corpus relies on publicly accessible documents and therefore cannot capture proprietary algorithms, internal moderation protocols, unpublished government requests, or users' direct experiences of enforcement. Documentary coding identifies patterned relationships without establishing causal effects or measuring platform-wide prevalence. Future research should combine policy analysis with interviews, platform audits, appeal records, computational trace data, and multilingual user studies across Indonesian regions. Longitudinal comparison is also needed to assess how SAMAN, generative-AI moderation, and localization policies evolve over time. Such work should test whether transparency, independent review, and local-language expertise materially reduce unequal visibility and procedural burden.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Reza Pahlawan: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly accessible regulatory, platform, and civil-society documents.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible regulatory, platform-policy, enforcement, and civil-society documents, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The corpus of documents analyzed in this study are publicly accessible and available from the corresponding author upon reasonable request.

REFERENCES

[1] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18647.

[2] Y. Mubarok, D. Sudana, D. Yanti, Sugiyo, A. Aisyah, and A. N. Af'idah, "Abusive Comments (Hate Speech) on Indonesian Social Media: A Forensic Linguistics Approach," *Theory and Practice in Language Studies*, May 2024, doi: 10.17507/tpls.1405.16.

[3] V. Gongane, M. Munot, and A. Anuse, "Detection and moderation of detrimental content on social media platforms: current status and future directions," *Social Network Analysis and Mining*, vol. 12, Sep. 2022, doi: 10.1007/s13278-022-00951-3.

[4] A. Fawaid, "Mapping the field of digital literary studies: Concepts, methods, and emerging directions," *Lingua Technica: Journal of Digital Literary Studies*, vol. 1, no. 1, pp. 1–13, Jan. 2025, doi: 10.64595/99c6t274.

[5] M. Singhal, C. Ling, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice," *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 868–895, Jun. 2022, doi: 10.1109/eurosp57164.2023.00056.

[6] R. Griffin, "Rethinking rights in social media governance: human rights, ideology and inequality," *European Law Open*, vol. 2, pp. 30–56, Mar. 2023, doi: 10.1017/elo.2023.7.

[7] D. Liu, "The Balancing Acts: Communicating Legitimacy in Global Speech Governance," *Social Media + Society*, vol. 11, Apr. 2025, doi: 10.1177/20563051251340855.

[8] O. Haimson, D. Delmonaco, P. Nie, and A. Wegner, "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, pp. 1–35, Oct. 2021, doi: 10.1145/3479610.

[9] B. Duffy and C. Meisner, "Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility," *Media, Culture & Society*, vol. 45, pp. 285–304, Jul. 2022, doi: 10.1177/01634437221111923.

[10] Y. Lyu, J. Cai, A. Callis, K. Cotter, and J. Carroll, "'I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.': A Study of Blind TikTokers' Content Moderation Experiences," presented at the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Jan. 2024. doi: 10.1145/3613904.3642148.

[11] S. Mayworm, M. Devito, D. Delmonaco, H. Thach, and O. Haimson, "Content Moderation Folk Theories and Perceptions of Platform Spirit among Marginalized Social Media Users," *ACM Transactions on Social Computing*, vol. 7, pp. 1–27, Dec. 2023, doi: 10.1145/3632741.

[12] H. Thach, S. Mayworm, D. Delmonaco, and O. Haimson, "(In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit," *New Media & Society*, vol. 26, pp. 4034–4055, Jul. 2022, doi: 10.1177/14614448221109804.

[13] A. ManishKumar and S. Gupta, "Governance of Social Media Platforms: A Literature Review," *Pac. Asia J. Assoc. Inf. Syst.*, vol. 15, p. 3, Mar. 2023, doi: 10.17705/1pais.15103.

[14] D. Endres, L. Hedler, and K. Wodajo, "Bias in Social Media Content Management: What Do Human Rights Have to Do with It?," *AJIL Unbound*, vol. 117, pp. 139–144, Jun. 2023, doi: 10.1017/aju.2023.23.

[15] S. Jhaver, A. Q. Zhang, Q. Z. Chen, N. Natarajan, R. Wang, and A. Zhang, "Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, pp. 1–33, May 2023, doi: 10.1145/3610080.

[16] L. Madio and M. Quinn, "Content Moderation and Advertising in Social Media Platforms," *SSRN Electronic Journal*, Jun. 2024, doi: 10.2139/ssrn.4875546.

[17] S. J. Kartutu and U. Rusadi, "Social Media as a New Arena of Power in Indonesia: A Political Economy Perspective in the Digital Platform Era," *International Journal of Social and Human*, Feb. 2025, doi: 10.59613/0pscjk39.

[18] B. Sander, "Democratic Disruption in the Age of Social Media: Between Marketized and Structural Conceptions of Human Rights Law," *European Journal of International Law*, Feb. 2021, doi: 10.1093/ejil/chab022.

[19] T. Divon, C. Are, and P. Briggs, "Platform gaslighting: A user-centric insight into social media corporate communications of content moderation," *Platforms & Society*, vol. 2, Jan. 2025, doi: 10.1177/29768624241303109.

[20] D. Oh and J. Downey, "Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI," *Information, Communication & Society*, vol. 28, pp. 1157–1176, May 2024, doi: 10.1080/1369118x.2024.2346531.

[21] R. Rogers, "Marginalizing the Mainstream: How Social Media Privilege Political Information," *Frontiers in Big Data*, vol. 4, Jul. 2021, doi: 10.3389/fdata.2021.689036.

[22] A. Heldt and S. Dreyer, "Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies From A Governance Structure Perspective," *Journal of Information Policy*, Jun. 2021, doi: 10.5325/jinfopoli.11.2021.0266.

[23] R. A. Wismashanti, "Social Media Content Moderation Challenges for Vulnerable Groups: A Case Study on Tiktok Indonesia," *Eduvest - Journal of Universal Studies*, Aug. 2023, doi: 10.59188/eduvest.v3i8.893.

[24] P. Pujiati, A. Lundeto, and I. Trianto, "Representing Arab-Indonesian identity: Language and cultural narratives on social media," *Indonesian Journal of Applied Linguistics*, Jan. 2025, doi: 10.17509/ijal.v14i3.78286.

[25] C. Peterson-Salahuddin, "Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation," *Big Data & Society*, vol. 11, Apr. 2024, doi: 10.1177/20539517241245333.
[26] S. Mayworm *et al.*, "The Online Identity Help Center: Designing and Developing a Content Moderation Policy Resource for Marginalized Social Media Users," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, pp. 1–30, Apr. 2024, doi: 10.1145/3637406.

BIOGRAPHIES OF AUTHORS

Reza Pahlawan is a lecturer and researcher at Universitas Negeri Yogyakarta, Indonesia. His academic concerns are language, design, communication, and related interdisciplinary fields. His research includes visual rhetoric, political communication, typography, multimodal discourse, electoral design, and media representation. He has been involved in academic discussions on how language and visual elements interact in shaping public meaning and political persuasion. He can be contacted at email: rezapahlawan@uny.ac.id