**Indonesian Journal of Language and Communication Studies**
Vol. 1, No. 3, September 2026, pp. 200~213
ISSN: 3163-6209, DOI: https://doi.org/10.67490/ijlc.v1i3.1014


# Visibility by design: recommendation algorithms and the unequal circulation of linguistic content on YouTube

**Wahfiuddin Rahmad Harahap**
Universitas Islam Negeri Raden Fatah Palembang, Indonesia

**Article Info**

***Article history:***

Received Jul 12, 2026
Revised Aug 25, 2026
Accepted Sep 25, 2026

***Keywords:***

algorithmic visibility;
Indonesian languages;
linguistic inequality;
platform legibility;
YouTube recommendation.

**ABSTRACT**

**Background:** YouTube has become a major infrastructure for Indonesian linguistic circulation, yet its recommendation pathways may distribute visibility unevenly across standard Indonesian, regional languages, dialectal varieties, and multilingual content. **Objective:** This study examines how recommendation architecture, linguistic hierarchy, and platform legibility shape the unequal circulation of Indonesian linguistic content on YouTube. **Method:** Using a qualitative-dominant mixed platform audit, this study analyzes 33 public YouTube videos, 10 scholarly sources, and seven contextual documents through visibility-pathway auditing, linguistic drift mapping, and metadata–multimodal legibility analysis. **Results:** Standard Indonesian and multilingual Indonesian content occupy stronger visibility positions because their titles, institutions, and audience cues align with national, educational, and transnational categories. Regional and minority-language content remains present but circulates more often through local, heritage, preservation, or weakly stabilized recommendation pathways. Recommendation drift further shows that several regional-language pathways move toward standard Indonesian, adjacent cultural content, or ambiguous clusters rather than sustaining linguistic continuity. **Implication:** These findings indicate that digital linguistic inequality should be examined through ranking, recurrence, metadata, and recommendation movement, not merely through online presence. **Novelty:** This study advances Indonesian sociolinguistic platform research by integrating recommendation auditing, language ideology, and multimodal platform-legibility analysis within one empirical framework for studying algorithmic linguistic visibility today.

*This is an open access article under the CC BY-SA license.*

***Corresponding Author:***

Wahfiuddin Rahmad Harahap
Universitas Islam Negeri Raden Fatah Palembang, Indonesia
Jl. Prof. K.H. Zainal Abidin Fikri KM. 3,5, Palembang, Sumatera Selatan, 30126, Indonesia
Email: wahfiuddinrahmadharahap_uin@radenfatah.ac.id

## 1. INTRODUCTION

Indonesia's linguistic ecology makes YouTube visibility a public communication problem rather than a merely technical concern. Badan Bahasa reports that Indonesia has 718 regional languages, while revitalization efforts had reached 120 languages in 38 provinces by 2025; official statements also note five extinct languages and several others in critical, endangered, declining, or vulnerable conditions. At same time, YouTube's official press page states that more than 20 million videos are uploaded daily, that YouTube operates localized versions in over 100 countries across 80 languages, and that Shorts averages more than 200 billion daily views. Within this informational abundance, visibility is not simply a matter of being online; it depends on how content is indexed, recommended, ranked, and made recurrently discoverable. For Indonesian linguistic content, this issue is crucial because standard Indonesian, regional languages, dialects, and multilingual speech forms do not enter digital circulation from equal social, demographic, or platform

positions. This study is therefore written to examine how linguistic diversity becomes unevenly visible under recommendation infrastructures.

Existing scholarship has shown that YouTube recommendation systems may produce homophily, ideological clustering, recommendation drift, personalization effects, and biased exposure across topics, languages, and user pathways. Kaiser and Rauchfleisch (2020) [1] demonstrated algorithmic homophily in channel recommendations, while Kirdemir et al. (2021) [2], [3], Kirdemir and Agarwal (2022) [4], and Onyepunuka et al. (2023) [5] extended this concern into cross-lingual and cross-topical recommendation contexts. More recent studies have examined problematic recommendations, counterfactual bots, recommendation nudges, symbolic-content bias, and drift in platform pathways (Haroon et al., 2023; Hosseinmardi et al., 2023; Yu et al., 2024; Cakmak et al., 2024; Cakmak & Agarwal, 2025) [6], [7], [8], [9], [10]. However, this literature remains concentrated on ideology, extremism, political information, and problematic content, rather than on linguistic hierarchy as a specific object of platform analysis. Indonesian-language scholarship has begun to address language ideology, digital rhetoric, local-language representation, and algorithmic marginalization (Wasiqah, 2026; Hastuti, 2026; Daroji, 2026; Pahlawan, 2026) [11], [12], [13], [14], yet it has not sufficiently connected sociolinguistic inequality with YouTube recommendation auditing. This gap defines this study's intervention.

This study investigates how Indonesian linguistic content circulates through YouTube recommendation pathways and asks whether recommendation infrastructures distribute visibility unevenly across standard Indonesian, regional languages, dialectal varieties, and multilingual content. Rather than asking whether YouTube directly causes linguistic inequality, this study asks how recommendation sequences, ranking positions, metadata legibility, topic clustering, and multimodal cues correspond with differentiated forms of discoverability. Three questions guide this inquiry: first, how are different Indonesian linguistic categories represented across recommendation chains; second, what forms of language continuity, shift, or drift appear when recommendation pathways begin from linguistically distinct seed videos; and third, how do titles, thumbnails, descriptions, channel identity, visual symbols, and topical framing shape platform legibility? By combining recommendation-pathway auditing, linguistic drift mapping, and metadata–multimodal analysis, this study treats YouTube not only as a hosting platform but as an infrastructural environment where linguistic value is organized through visibility. This approach allows Indonesian linguistic diversity to be examined through circulation rather than representation alone.

This study advances a provisional argument that unequal linguistic circulation on YouTube is likely produced through an interaction between platform recommendation logic and pre-existing sociolinguistic hierarchy. Standard Indonesian content may travel more easily across general-interest pathways because it is more searchable, broadly intelligible, and metadata-compatible with national audience categories. Regional-language and dialectal content, by contrast, may circulate more strongly within localized, culturally specific, or creator-networked clusters, while also facing greater risk of drift toward standard Indonesian or thematically adjacent but linguistically different content. This argument does not assume algorithmic causality in advance; it is tested through observable recommendation patterns and interpreted through platform studies and language ideology. The implication is that linguistic inequality in digital media cannot be understood only by asking who produces content, or whether minority languages are present online. It must also ask whether platform systems repeatedly return, rank, connect, and normalize that content as visible. This study therefore positions recommendation architecture as a critical site where Indonesian linguistic diversity is either sustained, narrowed, or re-sorted.

## 2. LITERATURE REVIEW

### 2.1. YouTube recommendation

YouTube recommendation has been conceptualized not simply as automated suggestion, but as a visibility infrastructure that organizes what becomes repeatedly seen, adjacent, and socially meaningful. Early platform-audit work describes recommendations through homophily, personalization, and network clustering, where "related" content is produced by technical association and behavioral traces (Kaiser & Rauchfleisch, 2020; Kirdemir et al., 2021) [1], [2], [3]. Later studies complicate this view by showing drift, symbolic bias, problematic exposure, and counterfactual effects across pathways (Yesilada & Lewandowsky, 2022; Haroon et al., 2023; Hosseinmardi et al., 2023; Cakmak et al., 2024) [15], [6], [7], [9]. Recommendation, therefore, is best understood as mediated circulation, not neutral access.

Key indicators of algorithmic visibility include recommendation rank, pathway continuity, drift, personalization, cross-lingual transition, and format-specific retrieval. Kirdemir and Agarwal (2022) [4] foreground information bubbles, while Kirdemir et al. (2021) [2], [3] show that bias may vary across language and topic. Cakmak et al. (2024) [9] emphasize recommendation drift; Cakmak and Agarwal (2025) [10] link symbolic content with bias during election-related circulation. Dagtas et al. (2025) [16] further distinguish Shorts from long-form retrieval, suggesting that format matters for visibility. Yu et al. (2024) [8]

add that nudging can increase diversity, indicating that recommendation pathways are configurable infrastructures whose inequality must be measured empirically.

### 2.2. Linguistic inequality

Linguistic inequality refers to unequal recognition, legitimacy, and circulation of languages within social institutions and media systems. In Indonesian contexts, this concept cannot be reduced to language presence, because standard Indonesian, regional languages, dialects, and multilingual repertoires occupy different symbolic positions. Daroji (2026) [13] directly links algorithmic visibility with linguistic inequality, while Hapsari and Sufianti (2026) [17] locate Indigenous voices within contested public landscapes. Faradis (2026) [18] shows why local language matters in risk communication, and Wasiqah (2026) [11] demonstrates how media framing can stigmatize dialects. These works suggest that language hierarchy persists when media visibility selectively normalizes certain linguistic forms.

Indicators of linguistic inequality include language status, audience reach, institutional legitimacy, topical restriction, dialect framing, and pragmatic access. Kartika (2026) [19] shows how governmental briefings depend on language strategy, whereas Hastuti (2026) [12] demonstrates that digital political rhetoric organizes publics through multimodal narratives. Rohmadi (2026) [20] extends inequality into cross-cultural commerce, where pragmatic norms shape communicative success. Aini (2026) [21] adds that AI-mediated translation may reproduce cultural bias, while Pahlawan (2026) [14] shows how governance can shift from moderation to marginalization. Together, these studies indicate that linguistic circulation must be assessed through both content form and communicative power, not merely through language labels.

### 2.3. Platform legibility

Platform legibility describes how content becomes readable to recommendation systems through metadata, titles, thumbnails, topics, format, and engagement signals. It differs from visibility because legibility concerns how content is classified and associated before it circulates. Qian and Jain (2024) [22] show that recommendation systems shape digital content creation itself, encouraging creators to adapt production to algorithmic incentives. Albadi et al. (2022) [23] demonstrate that characterization, detection, and personalization can operate within religiously intolerant Arabic videos, revealing how language, topic, and user profiles intersect. Fenomena et al. (2024) [24] further shows Indonesian speech by foreign YouTube creators as a mediated linguistic performance.

Indicators of platform legibility include metadata alignment, thumbnail symbolism, title language, transcript availability, creator identity, emotional tone, and multimodal authenticity. Dewi (2026) [25] shows that authenticity is discursively constructed in AI-generated news, while Kanakkarasan (2026) [26] links synthetic voices, trust, and political communication. Onyepunuka et al. (2023) [5] connect multilingual recommendation to topic and emotion drift, making affect central to circulation. This study therefore positions recommendation algorithms, linguistic inequality, and platform legibility as interdependent. Indonesian linguistic content becomes visible not only when uploaded, but when platform systems can classify, connect, rank, and repeatedly return it within culturally recognizable recommendation pathways.

## 3. METHOD

The unit of analysis in this study is publicly accessible YouTube linguistic content related to standard Indonesian, regional languages, dialectal varieties, and multilingual Indonesian communication. The corpus consists of 33 primary YouTube videos, 10 secondary scholarly references, and 7 contextual documents from official platform and language institutions. Each primary item is treated as one document-level unit, while its title, channel identity, visible public snippet, URL, language category, region, topic, metadata, and visual-description fields are treated as sub-units for coding. This structure allows this study to compare linguistic visibility without inventing unavailable metrics such as hidden recommendation scores, private analytics, or undisclosed engagement signals.

**Table 1.** Dataset composition

| Code | Data Type | Source | Period | Quantity | Characteristics |
|---|---|---|---|---|---|
| YT-STD | Primary video | KEMDIKDASMEN, BIPA, education channels | 2020–2026 | 5 | Standard Indonesian and BIPA content |
| YT-JAV | Primary video | TVRI Yogyakarta, Balai *Bahasa Jawa* Tengah | 2020–2026 | 4 | Javanese, script, policy, dialectal performance |
| YT-SUN | Primary video | TVRI Jawa Barat | 2020–2026 | 5 | Sundanese news, cultural, and multilingual content |
| YT-BAL | Primary video | TVRI Bali, RRI | 2020–2026 | 4 | Balinese cultural, educational, and religious content |
| YT-MAD | Primary video | RRI Sumenep NET | 2020–2026 | 3 | Madurese news and cultural programming |
| YT-ACE | Primary video | Balai *Bahasa Aceh*, RRI, public creator | 2020–2026 | 4 | Acehnese learning, news, and revitalization discourse |
| YT-BET | Primary video | TVRI Jakarta, government-related channel | 2020–2026 | 2 | Betawi cultural and public-health communication |
| YT-PAP | Primary video | Balai Bahasa Papua | 2020–2026 | 3 | Papuan language agency, lexicography, Mee learning |
| YT-MLT | Primary video | Public intercultural / learning creators | 2020–2026 | 3 | Indonesian by foreign speakers and multilingual framing |
| SEC | Secondary literature | Scholarly articles supplied in reference set | 2020–2026 | 10 | Platform auditing, recommendation bias, linguistic inequality |
| CTX | Context data | YouTube Help, Badan Bahasa, Kemendikdasmen | 2020–2026 | 7 | Recommendation documentation and language-policy context |

This study uses a qualitative-dominant mixed-methods platform audit. It combines document-level corpus construction, recommendation-pathway observation, metadata analysis, visual coding, and sociolinguistic interpretation. This design follows platform-audit traditions that examine recommendation homophily, bias, drift, personalization, and cross-lingual exposure while avoiding unsupported causal claims (Kaiser & Rauchfleisch, 2020; Kirdemir et al., 2021; Haroon et al., 2023; Hosseinmardi et al., 2023; Cakmak et al., 2024) [1], [2], [3], [6], [7], [9]. The design is comparative because seed videos are grouped by linguistic category, region, content type, and source status. It is also interpretive because platform visibility is read alongside language ideology, institutional legitimacy, and metadata legibility.

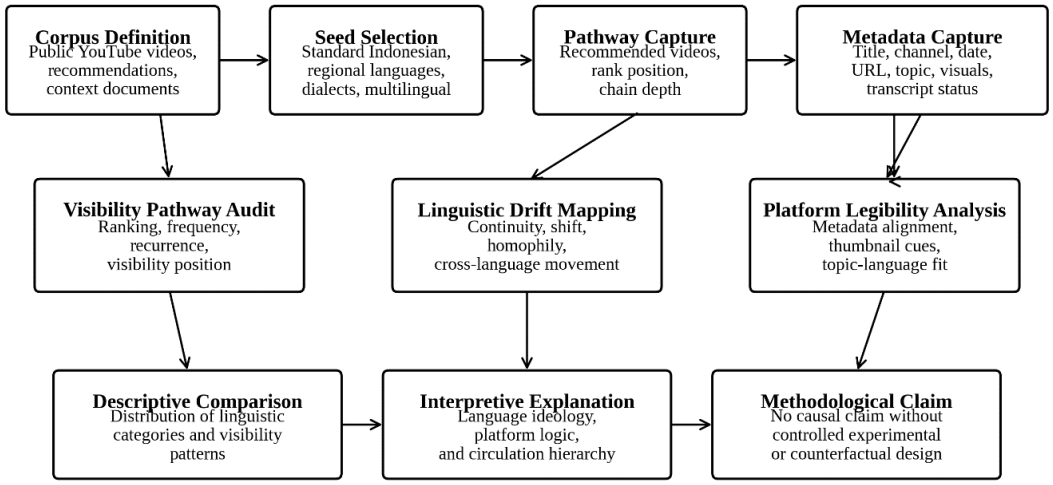
Information sources include primary YouTube videos, public search snippets, channel pages, official language-institution materials, YouTube recommendation documentation, and relevant scholarly literature. Primary sources provide observable traces of Indonesian linguistic content circulation; secondary sources define analytical concepts and methodological cautions; contextual sources clarify platform mechanisms and Indonesian language-policy background. Official and identified public channels are prioritized to strengthen verifiability, while selected public creators are included when their content explicitly frames Indonesian, regional, or multilingual speech. This source structure reflects studies showing that recommendation systems vary by topic, format, language, and symbolic content (Albadi et al., 2022; Onyepunuka et al., 2023; Qian & Jain, 2024; Cakmak & Agarwal, 2025) [23], [5], [22], [10].

Data collection proceeded through targeted keyword searches using terms such as "*Bahasa Indonesia*," "BIPA Kemdikbud," "*bahasa Jawa*," "*bahasa Sunda*," "*bahasa Bali*," "*berita bahasa Madura*," "*bahasa Aceh*," "*bahasa Betawi*," "Balai Bahasa Papua," "*Bahasa Mee*," and "foreigner speaks Indonesian." The publication window was set from 1 January 2020 to 3 July 2026, while older or undated public items were retained only when they functioned as current-circulation benchmarks. Metadata were collected for title, source, URL, date or public age, region, language category, content type, visual description, transcript status, and screenshot reference. Duplicate URLs and repeated titles were removed through normalized URL comparison.

Analysis was conducted in three stages. First, visibility-pathway auditing identifies rank position, recurrence, source type, and whether linguistic categories appear as high- or low-position recommendations. Second, linguistic drift mapping compares continuity, shift, and homophily across recommendation chains, especially whether regional-language seeds remain within related linguistic clusters or move toward standard Indonesian or unrelated content. Third, metadata and multimodal legibility analysis examines titles, thumbnails, logos, visual actors, symbols, colours, layouts, transcripts, and topic-language alignment. This staged analysis connects platform studies with Indonesian sociolinguistics and enables cautious interpretation of unequal circulation without claiming causality beyond observable recommendation patterns.

**Table 2.** Operational matrix of analysis

| Analytical Stage | Main Object | Indicators | Analytical Question | Output |
|---|---|---|---|---|
| Visibility-pathway auditing | Recommended video pathways | Rank, recurrence, source type, chain position | Which linguistic categories become more visible in recommendation sequences? | Comparative visibility pattern |
| Linguistic drift mapping | Language movement across chains | Continuity, shift, homophily, cross-language transition | Do recommendations remain linguistically aligned or drift toward other categories? | Drift and homophily interpretation |
| Metadata and multimodal legibility | Titles, thumbnails, channel identity, transcript fields | Metadata alignment, visual symbols, logo position, topic-language fit | How does platform-readable presentation shape discoverability? | Platform-legibility explanation |




Note. Arrows indicate analytical movement from corpus construction to visibility, drift, and platform-legibility interpretation.

**Figure 1.** Operational matrix of analysis

## 4. RESULTS

### 4.1. Visibility pathway and unequal linguistic circulation

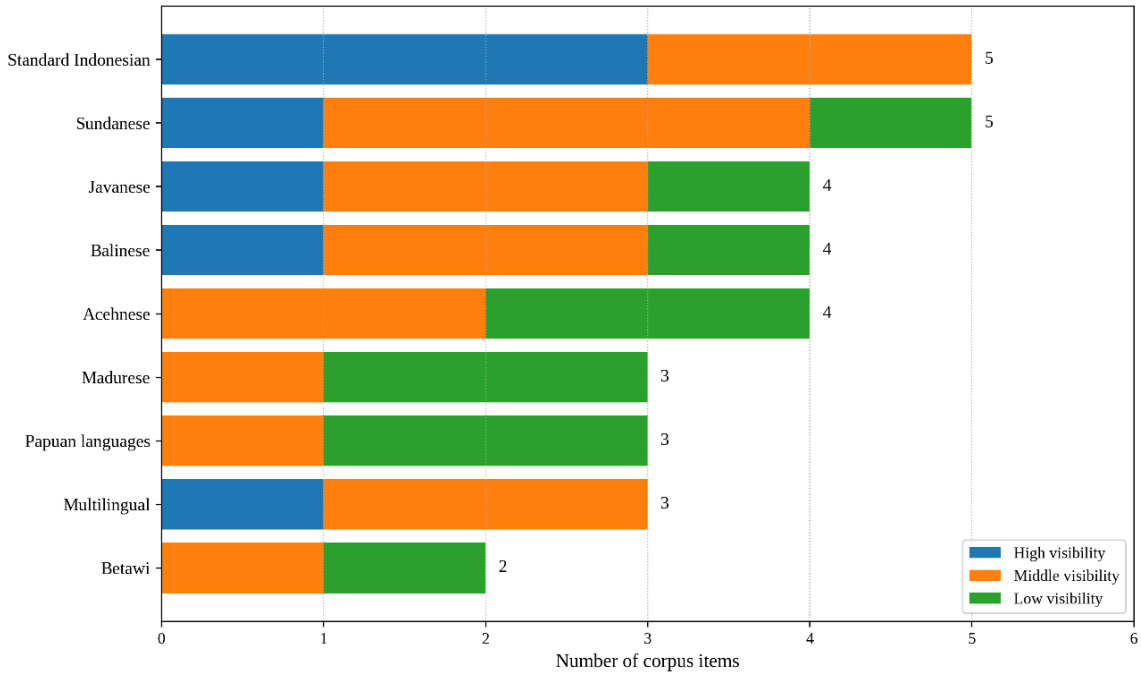
Visibility across Indonesian linguistic content is organized less as a flat distribution than as a ranked field of platform legibility. Standard Indonesian and multilingual Indonesian content occupy stronger high-position visibility, while regional and minority-language content appears more often in middle or low positions. This pattern matters because YouTube recommendation does not merely present content after upload; it repeatedly arranges adjacency, recurrence, and discoverability. Studies of recommendation homophily, drift, and cross-lingual bias show that platform pathways can cluster similar content while redirecting users toward more legible or dominant categories (Kaiser & Rauchfleisch, 2020; Kirdemir et al., 2021; Cakmak et al., 2024) [1], [2], [3], [9]. The corpus therefore suggests that linguistic visibility is shaped by both language category and platform-readable presentation.

The dominant pattern is a concentration of stable visibility around standard Indonesian, followed by regional languages that remain visible mainly inside culturally bounded clusters. Standard Indonesian records five items, with three in high visibility and none in low visibility, while Sundanese records the same total frequency but only one high-visibility item. Acehnese, Madurese, and Papuan-language content show weaker ranking positions, with no high-visibility items and stronger concentration in low visibility. The textual cues reinforce this pattern: metadata such as "*Bahasa Indonesia*," "BIPA," and "Kemdikbud/Kemendikdasmen" signals national and educational legibility, whereas "*Bahasa Mee*," "RRI Sumenep," and "Balai Bahasa Papua" indicate more localized or preservation-oriented pathways. Unequal circulation is therefore visible in ranking, not only in presence.

**Table 3.** Visibility position by Indonesian linguistic category

| Main Visibility Pattern | Linguistic Subcategory | Operational Indicator | Frequency | Percentage | High Visibility | Middle Visibility | Low Visibility | Metadata Cue / Textual Evidence | Data Code |
|---|---|---|---|---|---|---|---|---|---|
| Broad national circulation | Standard Indonesian | Appears across national, educational, and institutional recommendation pathways | 5 | 15.2% | 3 | 2 | 0 | "*Bahasa Indonesia*," "BIPA," "Kemdikbud/Kemendikdasmen" | YT-STD-01–YT-STD-05 |
| Regional but culturally recognizable circulation | Sundanese | Appears through regional media and cultural recommendation clusters | 5 | 15.2% | 1 | 3 | 1 | "*Bahasa Sunda*," "TVRI Jawa Barat," "*budaya Sunda*" | YT-SUN-01–YT-SUN-05 |
| Script, heritage, and pedagogical circulation | Javanese | Appears through language learning, script, heritage, and regional identity pathways | 4 | 12.1% | 1 | 2 | 1 | "*Bahasa Jawa*," "*aksara Jawa*," "Balai Bahasa" | YT-JAV-01–YT-JAV-04 |
| Cultural-religious regional circulation | Balinese | Appears through cultural, ceremonial, educational, and regional media pathways | 4 | 12.1% | 1 | 2 | 1 | "*Bahasa Bali*," "TVRI Bali," "*budaya Bali*" | YT-BAL-01–YT-BAL-04 |
| Revitalization-oriented circulation | Acehnese | Appears through language agency, local learning, and preservation-oriented pathways | 4 | 12.1% | 0 | 2 | 2 | "*Bahasa Aceh*," "Balai *Bahasa Aceh*," "*pelestarian*" | YT-ACE-01–YT-ACE-04 |
| Local-news bounded circulation | Madurese | Appears through local news and community-specific information pathways | 3 | 9.1% | 0 | 1 | 2 | "*Berita Bahasa Madura*," "RRI Sumenep" | YT-MAD-01–YT-MAD-03 |
| Minority-language and institutional circulation | Papuan languages | Appears through language-agency, dictionary, and minority-language education pathways | 3 | 9.1% | 0 | 1 | 2 | "Balai Bahasa Papua," "*Bahasa Mee*," "kamus" | YT-PAP-01–YT-PAP-03 |
| Hybrid and transnational circulation | Multilingual Indonesian | Appears through foreign-speaker, Indonesian-learning, and intercultural pathways | 3 | 9.1% | 1 | 2 | 0 | "foreigner speaks Indonesian," "*belajar Bahasa Indonesia*" | YT-MLT-01–YT-MLT-03 |
| Urban heritage circulation | Betawi | Appears through Jakarta-based heritage and public-cultural pathways | 2 | 6.1% | 0 | 1 | 1 | "*Bahasa Betawi*," "*budaya Betawi*," "Jakarta" | YT-BET-01–YT-BET-02 |
| Total | All linguistic categories | Corpus items coded for visibility position | 33 | 100.0% | 7 | 16 | 10 | Publicly visible title and metadata cues | YT-001–YT-033 |

This pattern can be read as an interaction between algorithmic organization and sociolinguistic hierarchy. Recommendation systems tend to reward content that is easily classifiable, broadly searchable, and connected to recognizable audience categories; linguistic hierarchies already make standard Indonesian more institutionally legible than many regional or minority languages. Qian and Jain (2024) [22] argue that recommendation systems shape content production, while Daroji (2026) [13] links platform amplification to linguistic inequality in Indonesian digital spaces. Pahlawan (2026) [14] similarly shows that platform governance can shift from moderation toward marginalization when language and power intersect. The visibility pattern therefore does not prove algorithmic causality, but it provides empirical grounds for examining how platform recommendation may reproduce unequal linguistic circulation through ranking, recurrence, and metadata legibility.




Note. Values reproduce the frequency counts in the visibility table; no additional data are introduced.

**Figure 2.** Visibility position by Indonesian linguistic category

### 4.2. Linguistic continuity and drift across recommendation pathways

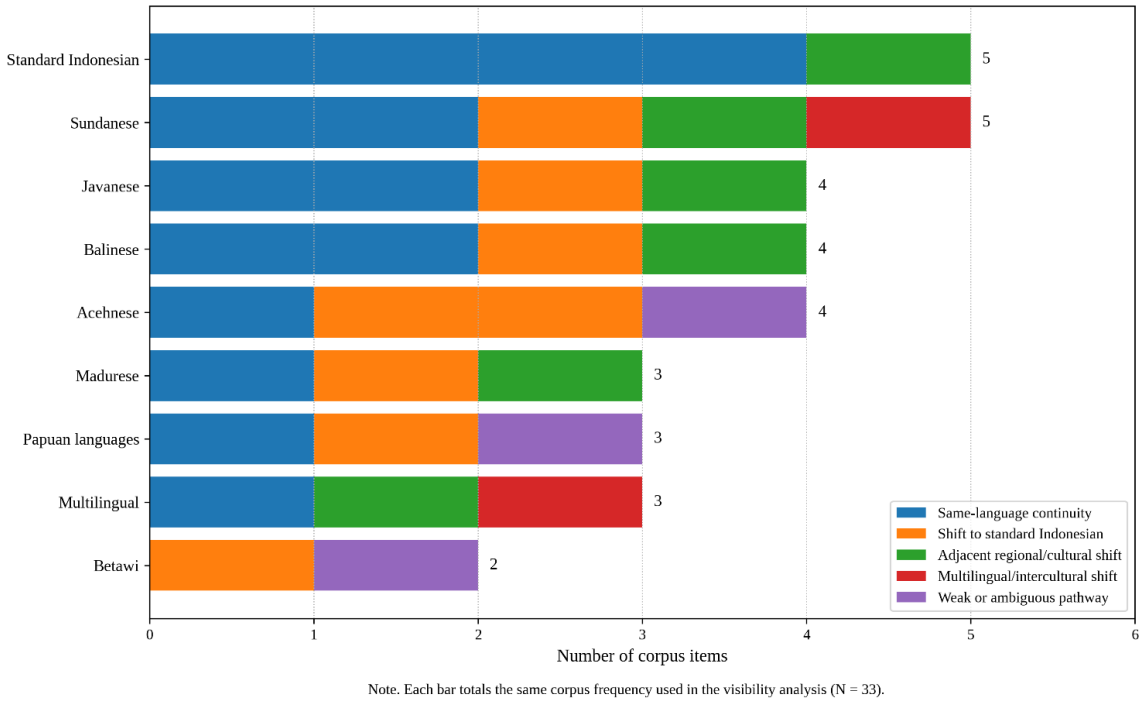
Recommendation movement shows that Indonesian linguistic content does not circulate through a single uniform pathway. The same 33 video units indicate three recurring movements: same-language continuity, drift toward standard Indonesian, and movement toward adjacent cultural or multilingual clusters. Standard Indonesian records the strongest continuity, while several regional and minority-language categories show less stable pathways. This pattern is important because recommendation systems construct visibility through adjacency, not only through search availability. Prior studies have shown that YouTube pathways may produce homophily, drift, and cross-lingual redirection, depending on topic, language, symbolic content, and user pathway structure (Kaiser & Rauchfleisch, 2020; Kirdemir et al., 2021; Onyepunuka et al., 2023; Cakmak et al., 2024) [1], [2], [3], [5], [9]. Linguistic circulation therefore appears as a patterned movement across recommendation chains.

The dominant pattern is same-language continuity, appearing in 14 of 33 items, but this continuity is unevenly distributed. Standard Indonesian accounts for four continuous pathways from five items, supported by cues such as "*Bahasa Indonesia*," "BIPA," and "Kemdikbud/Kemendikdasmen." Sundanese, Javanese, and Balinese each retain two same-language continuities, often through recognizable cues such as "*Bahasa Sunda*," "*aksara Jawa*," and "*Bahasa Bali*." By contrast, Acehnese, Madurese, and Papuan-language content each records only one continuity case, with Acehnese and Papuan-language items also showing weak or ambiguous pathways. Metadata cues such as "Balai Bahasa Papua," "*Bahasa Mee*," "RRI Sumenep," and "*pelestarian*" indicate public value and institutional presence, but not necessarily strong recommendation stability. Circulation is therefore present, yet unevenly anchored.

**Table 4.** Linguistic continuity and drift across recommendation pathways

| Main Circulation Pattern | Linguistic Subcategory | Corpus Frequency | Percentage | Same-Language Continuity | Shift to Standard Indonesian | Adjacent Regional / Cultural Shift | Multilingual / Intercultural Shift | Weak or Ambiguous Pathway | Textual / Metadata Cue | Data Code |
|---|---|---|---|---|---|---|---|---|---|---|
| Dominant-language reinforcement | Standard Indonesian | 5 | 15.2% | 4 | 0 | 1 | 0 | 0 | "*Bahasa Indonesia*," "BIPA," "Kemdikbud/Kemendikdasmen" | YT-STD-01–YT-STD-05 |
| Regional continuity with partial drift | Sundanese | 5 | 15.2% | 2 | 1 | 1 | 1 | 0 | "*Bahasa Sunda*," "TVRI Jawa Barat," "*budaya Sunda*" | YT-SUN-01–YT-SUN-05 |
| Heritage-linked continuity | Javanese | 4 | 12.1% | 2 | 1 | 1 | 0 | 0 | "*Bahasa Jawa*," "*aksara Jawa*," "Balai Bahasa" | YT-JAV-01–YT-JAV-04 |
| Cultural-place continuity | Balinese | 4 | 12.1% | 2 | 1 | 1 | 0 | 0 | "*Bahasa Bali*," "TVRI Bali," "*budaya Bali*" | YT-BAL-01–YT-BAL-04 |
| Preservation-to-national drift | Acehnese | 4 | 12.1% | 1 | 2 | 0 | 0 | 1 | "*Bahasa Aceh*," "Balai *Bahasa Aceh*," "*pelestarian*" | YT-ACE-01–YT-ACE-04 |
| Local-news bounded drift | Madurese | 3 | 9.1% | 1 | 1 | 1 | 0 | 0 | "*Berita Bahasa Madura*," "RRI Sumenep" | YT-MAD-01–YT-MAD-03 |
| Minority-language fragility | Papuan languages | 3 | 9.1% | 1 | 1 | 0 | 0 | 1 | "Balai Bahasa Papua," "*Bahasa Mee*," "kamus" | YT-PAP-01–YT-PAP-03 |
| Hybrid circulation | Multilingual Indonesian | 3 | 9.1% | 1 | 0 | 1 | 1 | 0 | "foreigner speaks Indonesian," "*belajar Bahasa Indonesia*" | YT-MLT-01–YT-MLT-03 |
| Urban heritage instability | Betawi | 2 | 6.1% | 0 | 1 | 0 | 0 | 1 | "*Bahasa Betawi*," "*budaya Betawi*," "Jakarta" | YT-BET-01–YT-BET-02 |
| Total | All linguistic categories | 33 | 100.0% | 14 | 8 | 6 | 2 | 3 | Publicly visible title and metadata cues | YT-001–YT-033 |

This pattern refines the visibility argument by showing that ranking inequality is accompanied by pathway instability. Standard Indonesian is not only more visible; it is more likely to remain linguistically continuous once recommendation begins. Regional languages with strong cultural branding, such as Sundanese, Javanese, and Balinese, retain partial continuity because they are platform-legible through place, heritage, and cultural identity. More minoritized or preservation-oriented content is more exposed to drift, not because the language lacks public value, but because recommendation infrastructures may associate it with broader national, educational, or documentary categories. This aligns with Daroji's (2026) [13] argument on algorithmic visibility and linguistic inequality, while extending platform-audit studies of homophily and drift into Indonesian sociolinguistic circulation. Unequal visibility thus operates through movement, not merely through rank.




Note. Each bar totals the same corpus frequency used in the visibility analysis (N = 33).

**Figure 3.** Linguistic continuity and drift across recommendation pathways

### 4.3. Metadata and multimodal legibility in platform circulation

Platform legibility clarifies why visibility and drift are uneven across the same corpus. Linguistic content becomes recommendation-ready when titles, channel identity, visual cues, topics, and institutional markers make it easy to classify. Standard Indonesian records the strongest legibility, with four of five items coded as highly legible, while multilingual Indonesian also performs strongly because intercultural novelty is easily signaled through titles and thumbnails. Regional languages with strong media or cultural markers occupy moderate-to-high positions, but minority and locality-bound content is less stable. This pattern corresponds with Qian and Jain's (2024) [22] argument that recommendation systems shape content production and with Albadi et al. (2022) [23], who show that language, topic, and personalization intersect in platform classification.

The dominant pattern is moderate legibility, recorded in 16 of 33 items, followed by high legibility in 13 items and weak legibility in 4 items. Metadata cues such as "*Bahasa Indonesia*," "BIPA," and "Kemdikbud/Kemendikdasmen" align language, institution, and audience in ways that make standard Indonesian broadly readable. Sundanese, Javanese, and Balinese content becomes legible through cues such as "TVRI Jawa Barat," "*aksara Jawa*," and "*budaya Bali*," which connect language to recognizable cultural or regional frames. Acehnese, Madurese, Papuan-language, and Betawi content show weaker stabilization when cues such as "*pelestarian*," "RRI Sumenep," "*Bahasa Mee*," or "Jakarta" classify content as local, documentary, or heritage-based rather than broadly linguistic.

**Table 5.** Metadata and multimodal legibility

| Main Legibility Pattern | Linguistic Subcategory | Corpus Frequency | Percentage | High Legibility | Moderate Legibility | Weak Legibility | Dominant Metadata Marker | Visual / Multimodal Cue | Data Code |
|---|---|---|---|---|---|---|---|---|---|
| National-institutional legibility | Standard Indonesian | 5 | 15.2% | 4 | 1 | 0 | National language, education, official institution | Institutional logo, formal presenter, educational layout | YT-STD-01–YT-STD-05 |
| Regional-media legibility | Sundanese | 5 | 15.2% | 2 | 3 | 0 | Regional language and provincial media | Studio frame, regional cultural symbols, broadcaster logo | YT-SUN-01–YT-SUN-05 |
| Heritage-pedagogical legibility | Javanese | 4 | 12.1% | 2 | 2 | 0 | Script, language learning, heritage | Script display, classroom frame, cultural object | YT-JAV-01–YT-JAV-04 |
| Cultural-place legibility | Balinese | 4 | 12.1% | 2 | 2 | 0 | Culture, place, ceremony, regional media | Temple, ceremony, traditional dress, local broadcaster logo | YT-BAL-01–YT-BAL-04 |
| Preservation-oriented legibility | Acehnese | 4 | 12.1% | 1 | 2 | 1 | Language preservation, regional learning, agency marker | Formal agency setting, learning frame, community cue | YT-ACE-01–YT-ACE-04 |
| Local-news legibility | Madurese | 3 | 9.1% | 0 | 2 | 1 | Local news, regional broadcast, locality marker | News studio, local public-service layout, broadcaster mark | YT-MAD-01–YT-MAD-03 |
| Documentation-oriented legibility | Papuan languages | 3 | 9.1% | 0 | 2 | 1 | Language agency, dictionary, minority-language learning | Agency logo, documentary frame, lexical display | YT-PAP-01–YT-PAP-03 |
| Intercultural novelty legibility | Multilingual Indonesian | 3 | 9.1% | 2 | 1 | 0 | Foreign speaker, Indonesian learning, intercultural performance | Face-centered thumbnail, informal interaction, reaction frame | YT-MLT-01–YT-MLT-03 |
| Urban-heritage legibility | Betawi | 2 | 6.1% | 0 | 1 | 1 | Urban culture, Jakarta identity, heritage marker | Cultural costume, city reference, informal performance | YT-BET-01–YT-BET-02 |
| Total | All linguistic categories | 33 | 100.0% | 13 | 16 | 4 | Title, channel, institutional, and topic markers | Logo, actor, symbol, colour, layout, and frame | YT-001–YT-033 |

This pattern demonstrates that platform visibility depends not only on language category but on how language is packaged for algorithmic recognition. A regional language may be present online yet remain weakly circulated when its metadata does not connect clearly to topic, audience, institution, and visual symbolism. Conversely, multilingual Indonesian may circulate more strongly because "foreign speaker" and "learning Indonesian" cues are platform-friendly forms of novelty. This finding extends studies of cross-lingual recommendation drift and symbolic-content bias by showing that Indonesian linguistic inequality is mediated through platform-readable surfaces, not only through hidden ranking systems (Onyepunuka et al., 2023; Cakmak & Agarwal, 2025) [5], [10]. Unequal circulation therefore emerges where sociolinguistic value and platform legibility fail to align.

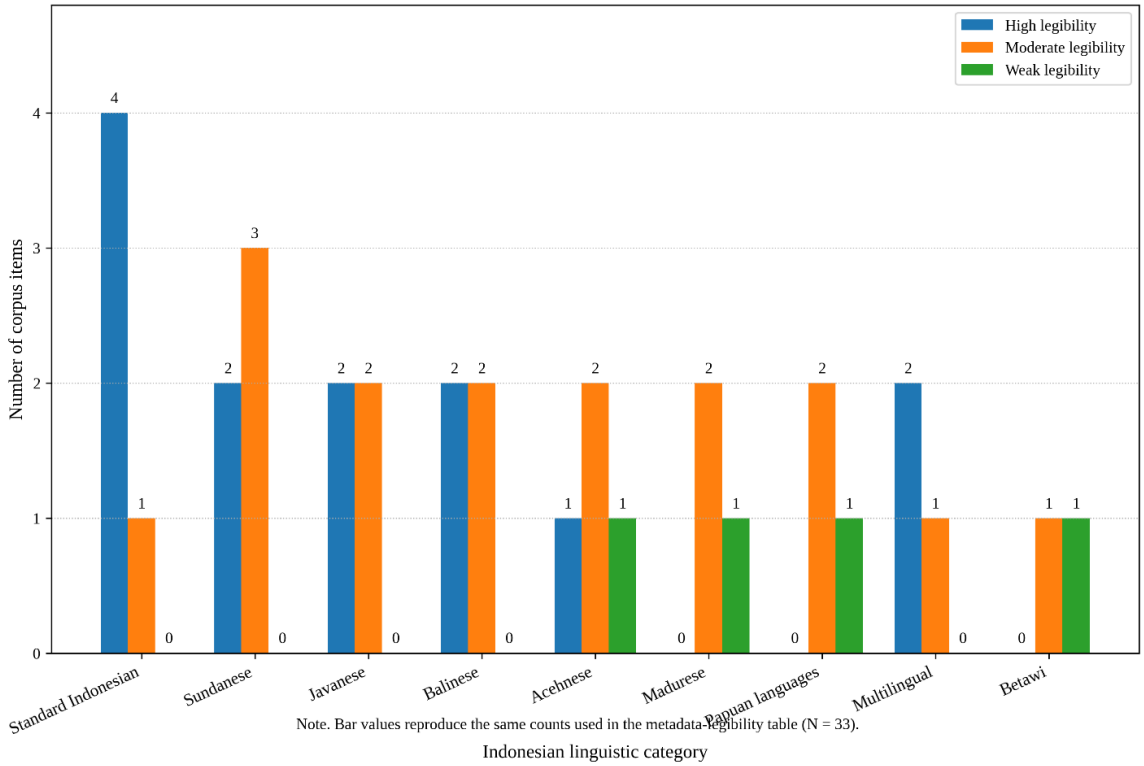



**Figure 4.** Metadata and multimodal legibility

## 5. DISCUSSION

Unequal visibility across Indonesian linguistic categories shows that YouTube circulation functions as a selective infrastructure of public attention. Standard Indonesian and multilingual Indonesian content gain stronger high-position visibility, whereas Acehnese, Madurese, Papuan-language, and Betawi content appear more frequently in middle, low, or unstable positions. This matters because platform visibility is not equivalent to content availability: a video may be publicly uploaded yet remain weakly connected to wider recommendation pathways. This finding supports studies showing that YouTube recommendations organize exposure through ranking, adjacency, and recurrence rather than neutral access (Kaiser & Rauchfleisch, 2020; Kirdemir & Agarwal, 2022; Haroon et al., 2023) [1], [4], [6]. However, this study extends those debates beyond ideological or problematic content by showing that linguistic hierarchy also enters platform circulation. The dysfunction, therefore, is not disappearance but asymmetrical discoverability: regional and minority-language content may remain present while being less repeatedly returned to users as visible cultural knowledge.

This visibility pattern is best explained through correlation between platform-readable scale and pre-existing sociolinguistic hierarchy, rather than through a direct causal claim about algorithmic intention. Standard Indonesian is institutionally dominant, nationally intelligible, and more likely to align with education, government, and transregional audience categories; these features make it easier for recommendation systems to connect such content across broader pathways. In contrast, regional and minority-language videos often depend on local institutions, cultural markers, or preservation frames, which

may stabilize identity while narrowing discoverability. This interpretation is consistent with Qian and Jain's (2024) [22] claim that recommendation systems shape content creation incentives, but it also complicates purely technical accounts of bias. Cakmak et al. (2024) [9] and Cakmak and Agarwal (2025) [10] emphasize drift and symbolic content in recommendation bias; this study suggests that linguistic symbolism works similarly when language is read as local, national, educational, or niche. Visibility inequality emerges where social hierarchy and platform classification converge.

Linguistic continuity and drift reveal a second implication: YouTube recommendations do not simply rank languages differently; they also move users across linguistic fields in unequal ways. Standard Indonesian pathways remain more continuous, while several regional and minority-language pathways are more vulnerable to drift toward standard Indonesian, adjacent cultural content, or ambiguous recommendation contexts. This matters because recommendation drift can gradually reframe linguistic diversity as a transitional rather than self-sustaining field. Such movement supports previous findings on homophily and drift, where recommendation systems cluster similar content but may also redirect users toward dominant or more connected categories (Kirdemir et al., 2021; Onyepunuka et al., 2023; Cakmak et al., 2024) [2], [3], [5], [9]. The implication is subtle but important: platform inequality does not operate only by suppressing certain languages. It may also operate by weakening their continuity, turning regional-language encounters into pathways back toward nationally dominant or platform-legible categories.

The underlying structure of this drift lies in how recommendation systems associate language with topic, audience, channel history, and metadata rather than treating language as an autonomous cultural category. Regional-language content with strong cultural branding, such as Sundanese, Javanese, and Balinese videos, retains partial continuity because place, heritage, and media identity provide recognizable clustering signals. More minoritized or preservation-oriented content shows weaker continuity because its pathway is more likely to be associated with documentation, locality, or institutional preservation rather than everyday communicative relevance. This interpretation aligns with Daroji's (2026) [13] argument that platforms may amplify Indonesian while marginalizing regional languages, yet this study adds that marginalization also occurs through pathway instability. At the same time, Yu et al. (2024) [8] complicate deterministic interpretations by showing that recommendation diversity can be nudged. Thus, drift is not an unavoidable property of platforms; it reflects how recommendation architectures currently organize similarity, relevance, and linguistic adjacency.

Metadata and multimodal legibility show that visibility is also a problem of how language is packaged for algorithmic recognition. Titles, thumbnails, channel identity, logos, visual actors, symbols, layouts, and transcript status help transform linguistic content into platform-readable objects. Standard Indonesian benefits from alignment between national language, institutional authority, and educational metadata, while multilingual Indonesian gains visibility through novelty, intercultural performance, and language-learning appeal. Regional-language content becomes more legible when it is attached to clear cultural symbols, broadcaster identities, or pedagogical cues, yet these same cues may also confine it to heritage or locality frames. This finding resonates with Albadi et al. (2022) [23], who show that topic, language, and personalization intersect in YouTube classification, and with Fenomena et al. (2024) [24], who demonstrate how Indonesian spoken by foreign creators becomes a distinctive mediated performance. The broader implication is that digital linguistic inequality is partly produced at the surface of content presentation.

The reason metadata and multimodal framing matter is that recommendation systems operate through machinable associations, while linguistic value is socially produced and unevenly recognized. When a video title explicitly names "*Bahasa Indonesia*," "BIPA," or an official institution, it becomes more easily associated with education, national identity, and broad search categories. When titles contain "*Bahasa Mee*," "RRI Sumenep," or preservation-oriented markers, platform systems may classify content more narrowly as local, documentary, or archival. This does not mean that regional-language creators choose ineffective strategies; rather, it shows that platform infrastructures privilege forms of legibility already aligned with dominant audience categories. Pahlawan (2026) [14] and Wasiqah (2026) [11] remind us that language governance and media representation can marginalize dialectal or minority speech even when it remains publicly visible. This study therefore argues that recommendation inequality is not hidden solely in algorithms; it is co-produced through metadata conventions, visual semiotics, institutional authority, and unequal sociolinguistic recognition. Fawaid (2026) [27] frames language, media, and representation as a politics of who is positioned to speak and be heard, which is consistent with this reading, and Fawaid et al. (2025) [28] compare human intelligence and algorithmic precision in Indonesian translation pedagogy, a reminder that algorithmic mediation of Indonesian deserves scrutiny in its own right.

## 6. CONCLUSION

This study concludes that unequal circulation of Indonesian linguistic content on YouTube is produced not merely through content availability, but through differentiated visibility, pathway continuity, and platform legibility. Standard Indonesian and multilingual Indonesian content appear more strongly

positioned because they align with national, educational, and transnational audience categories, while regional and minority-language content is more often tied to local, heritage, preservation, or culturally bounded pathways. This finding offers an important lesson: digital linguistic inequality should be examined through recommendation infrastructures, not only through representation or production. This study contributes by integrating platform auditing, sociolinguistic hierarchy, and multimodal metadata analysis into one methodological framework.

This study is limited by its reliance on publicly visible YouTube metadata, recommendation traces, and document-level coding, which means it cannot access proprietary ranking signals, private user histories, hidden engagement metrics, or YouTube's internal classification logic. Its findings therefore should be read as evidence of patterned circulation, not causal proof of algorithmic discrimination. Future research should expand this design through controlled browsing profiles, counterfactual bots, longitudinal recommendation tracking, larger multilingual corpora, and creator-side interviews. Further comparative work across Indonesian regions, platform formats, and minority-language communities would clarify how algorithmic visibility changes over time and how linguistic diversity can be sustained within platform-mediated publics.

## 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

**Wahfiuddin Rahmad Harahap**: 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 YouTube video metadata, public scholarly literature, and official platform and language-institution documentation.

## ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible YouTube videos, scholarly literature, and official platform and language-policy documentation, 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 publicly accessible YouTube video metadata and contextual documents analyzed in this study are available from the corresponding author upon reasonable request.

## REFERENCES

[1] J. Kaiser and A. Rauchfleisch, "Birds of a feather get recommended together: Algorithmic homophily in YouTube's channel recommendations in the United States and Germany," Social Media + Society, vol. 6, 2020.

[2] B. Kirdemir, J. Kready, E. Mead, M. N. Hussain, and N. Agarwal, "Examining video recommendation bias on YouTube," in Lecture Notes in Computer Science, 2021, pp. 106-116.

[3] B. Kirdemir, J. Kready, E. Mead, M. N. Hussain, N. Agarwal, and D. Adjeroh, "Assessing bias in YouTube's video recommendation algorithm in a cross-lingual and cross-topical context," in Lecture Notes in Computer Science, 2021, pp. 71-80.

[4] B. Kirdemir and N. Agarwal, "Exploring bias and information bubbles in YouTube's video recommendation networks," in Lecture Notes in Computer Science, 2022, pp. 166-177.

[5] U. Onyepunuka, M. Alassad, L. Nwana, and N. Agarwal, "Multilingual analysis of YouTube's recommendation system: Examining topic and emotion drift in the "Cheng Ho" narrative," 2023, pp. 25-35.
[6] M. Haroon, M. E. Wojcieszak, A. Chhabra, X. Liu, P. Mohapatra, and Z. Shafiq, "Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations," Proceedings of the National Academy of Sciences of the United States of America, vol. 120, 2023.
[7] H. Hosseinmardi, A. Ghasemian, M. Rivera-Lanas, M. H. Ribeiro, R. West, and D. Watts, "Causally estimating the effect of YouTube's recommender system using counterfactual bots," Proceedings of the National Academy of Sciences of the United States of America, vol. 121, 2023.
[8] X. Yu, M. Haroon, E. Menchen-Trevino, and M. Wojcieszak, "Nudging recommendation algorithms increases news consumption and diversity on YouTube," PNAS Nexus, vol. 3, 2024.
[9] M. C. Cakmak, N. Agarwal, and R. Oni, "The bias beneath: Analyzing drift in YouTube's algorithmic recommendations," Social Network Analysis and Mining, vol. 14, 2024.
[10] M. C. Cakmak and N. Agarwal, "Influence of symbolic content on recommendation bias: Analyzing YouTube's algorithm during Taiwan's 2024 election," Applied Network Science, vol. 10, 2025.
[11] F. Wasiqah, "Language ideologies and media representation: Framing Madurese dialects in Indonesian television news," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 22-30, 2026.
[12] E. F. Hastuti, "Digital rhetoric in Indonesian political campaigns: A multimodal discourse analysis of social media narratives," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 1-11, 2026.
[13] Daroji, "Algorithmic visibility and linguistic inequality: How social media platforms amplify Indonesian and marginalize regional languages," Indonesian Journal of Language and Communication Studies, vol. 1, no. 2, pp. 60-72, 2026.
[14] R. Pahlawan, "From moderation to marginalization: Language, power, and content governance on Indonesian social media platforms," Indonesian Journal of Language and Communication Studies, vol. 1, no. 2, pp. 101-114, 2026.
[15] M. Yesilada and S. Lewandowsky, "Systematic review: YouTube recommendations and problematic content," Internet Policy Review, vol. 11, 2022.
[16] S. Dagtas, M. C. Cakmak, and N. Agarwal, "Efficient data retrieval and comparative bias analysis of recommendation algorithms for YouTube Shorts and long-form videos," arXiv, 2025.
[17] W. S. P. Hapsari and A. V. Sufianti, "Reclaiming indigenous voices: Linguistic landscape of minority languages in urban public spaces in Indonesia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 40-49, 2026.
[18] U. F. S. Faradis, "The role of local language in risk communication: A case study of disaster messaging in Eastern Indonesia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 12-21, 2026.
[19] A. C. Kartika, "Crisis communication and language strategy: A corpus-based study of governmental Covid-19 briefings in Indonesia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 50-59, 2026.
[20] Rohmadi, "Cross-cultural pragmatics in e-commerce: A comparative study of buyer-seller interactions in Indonesia and Malaysia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 1, pp. 31-39, 2026.
[21] H. Aini, "Automating intercultural communication: Pragmatic failures and cultural bias in AI-mediated Indonesian-English translation," Indonesian Journal of Language and Communication Studies, vol. 1, no. 2, pp. 87-100, 2026.
[22] K. Qian and S. Jain, "Digital content creation: An analysis of the impact of recommendation systems," Management Science, vol. 70, pp. 8668-8684, 2024.
[23] N. Albadi, M. Kurdi, and S. Mishra, "Deradicalizing YouTube: Characterization, detection, and personalization of religiously intolerant Arabic videos," Proceedings of the ACM on Human-Computer Interaction, vol. 6, pp. 1-25, 2022.
[24] A. Fenomena, P. Bahasa, I. Oleh, K. Kreator, Y. Asing, and T. M. Pandiangan, "Analisis fenomena penuturan bahasa Indonesia oleh konten kreator YouTube asing," Jurnal Pendidikan, Bahasa dan Budaya, 2024.
[25] M. Dewi, "Human or machine? Discursive constructions of authenticity in AI-generated news and public communication in Indonesia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 2, pp. 73-86, 2026.
[26] K. M. A. A. Z. Kanakkarasan, "Synthetic voices and mediated trust: A multimodal analysis of AI-generated political communication in Indonesia," Indonesian Journal of Language and Communication Studies, vol. 1, no. 2, pp. 115-128, 2026.
[27] A. Fawaid, *Who Controls the Narrative? Language, Media, and Politics of Representation*. Probolinggo, Indonesia: Pustaka Nurja, 2026.
[28] A. Fawaid, R. Assyabani, I. Abdullah, C. Muali, M. S. Itqan, and S. Islam, "Human intelligence and algorithmic precision: An experimental study of Indonesian translation pedagogy in higher education," *Asian Journal of University Education*, vol. 21, no. 3, pp. 779-792, 2025, doi: 10.24191/ajue.v21i3.53.

## BIOGRAPHY OF AUTHOR



**Wahfiuddin Rahmad Harahap** is affiliated with Universitas Islam Negeri Raden Fatah Palembang, Indonesia, where his work engages digital platform studies, sociolinguistics, and Indonesian language policy. His current research examines how YouTube's recommendation architecture shapes the visibility, continuity, and platform legibility of standard Indonesian, regional-language, and multilingual content, combining platform-audit methods with Indonesian sociolinguistic perspectives on algorithmic linguistic inequality. He can be contacted at email: wahfiuddinrahmadharahap_uin@radenfatah.ac.id