

Algorithmic visibility and linguistic inequality: how social media platforms amplify Indonesian and marginalize regional languages

Daroji

Universitas Sarjanawiyata Tamansiswa, Yogyakarta, Indonesia

Article Info

Article history:

Received Apr 23, 2026

Revised May 25, 2026

Accepted Jun 30, 2026

Keywords:

algorithmic visibility;
linguistic inequality;
regional languages;
YouTube.

ABSTRACT

Background: Social media platforms increasingly mediate linguistic visibility, yet numerical presence alone cannot reveal whether regional languages receive comparable exposure, functional diversity, or infrastructural support within multilingual digital environments. **Objective:** This study examines how YouTube structures the visibility of Indonesian, Javanese, and Madurese content through differences in retrieval, thematic distribution, and dominant-language mediation. **Method:** Using a sequential mixed-method platform-audit design, the study combines a 30-video discovery corpus with controlled visibility auditing, stratified quantitative content analysis, and sociolinguistic domain analysis across six predefined thematic categories. **Results:** The corpus was numerically balanced across the three languages, but its thematic composition and verification readiness differed substantially. Javanese covered all six thematic categories, Indonesian five, and Madurese four, with Madurese candidates concentrated in humour and news. Linguistic evidence was strongest for Madurese, while Indonesian was frequently inferred as an unmarked default and regional-language content sometimes relied on Indonesian subtitles or translations for wider accessibility. **Implication:** The findings suggest that platform-based linguistic inequality is based on the interaction of ranked exposure, functional-domain breadth, evidential certainty, and dominant-language mediation rather than as a simple disparity in content counts, thereby. **Novelty:** The novelty of this study is a replicable framework for examining multilingual visibility without attributing observed differences prematurely to intentional algorithmic suppression at platform scale.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Daroji

Universitas Sarjanawiyata Tamansiswa, Yogyakarta, Indonesia

Jl. Batikan UH III/1043, Tahunan, Umbulharjo, Kota Yogyakarta, 55167, Indonesia

Email: daroji250@gmail.com

1. INTRODUCTION

Indonesia's linguistic diversity is increasingly mediated by digital platforms whose ranking and recommendation systems determine which languages become discoverable, repeatable, and economically valuable. The country has 718 living languages, yet UNESCO notes that most possess little substantial online presence compared with Indonesian, the national language. This imbalance matters in a society where internet use has become structurally embedded: the Indonesian Internet Service Providers Association reported 221.56 million internet users in 2024, equivalent to 79.5% of the population. Social media platforms therefore operate not merely as communication channels but as infrastructures through which linguistic legitimacy and public attention are distributed. Although regional languages such as Javanese and Madurese

remain important in familial, cultural, and community domains, their social vitality does not automatically secure equivalent visibility online. When search and recommendation systems privilege content associated with larger audiences, richer metadata, or stronger engagement histories, digital participation may reproduce rather than diminish existing linguistic hierarchies [1], [2].

Existing scholarship has established that algorithmic systems sort, rank, and selectively expose content, thereby shaping the conditions under which users and creators become visible. Research on Instagram, TikTok, and other platforms has shown that creators must negotiate opaque and continually changing “visibility games,” while algorithmic curation can unevenly distribute attention across actors, genres, and communicative practices [3], [4]. Parallel research in language technology demonstrates systematic performance disparities between extensively resourced languages and the majority of the world’s languages, whose smaller digital footprints produce weaker technological support [5], [6], [7]. Sociolinguistic studies have likewise examined digital multilingualism, minority-language maintenance, and the ideological regimentation of language through commercial technologies [8], [9]. Nevertheless, these strands remain insufficiently connected. Research rarely investigates how platform ranking intersects with internal linguistic inequality within highly multilingual nation-states. More specifically, little empirical work compares the algorithmic visibility, thematic distribution, and Indonesian-language mediation of regional-language videos on YouTube.

This study investigates how YouTube’s platform environment structures the comparative visibility of Indonesian, Javanese, and Madurese content. It addresses three interrelated questions. First, under controlled and semantically comparable search conditions, how are videos using the three languages distributed across ranked search positions? Second, does the thematic breadth of retrieved content vary by language, particularly across education, humour, food, local culture, news, and everyday life? Third, to what extent is the digital visibility of Javanese and Madurese mediated through Indonesian-language titles, descriptions, subtitles, or mixed-language presentation? These questions distinguish ranked exposure from numerical availability and linguistic presence from cultural reference. A video concerning Javanese food, for example, cannot automatically be treated as Javanese-language content unless its spoken language and metadata are verified. By combining a controlled algorithmic visibility audit, stratified quantitative content analysis, and sociolinguistic domain analysis, the study examines visibility as a multidimensional condition encompassing ranking, thematic functionality, metadata congruence, and translational mediation rather than as a simple count of available videos.

The study advances the proposition that linguistic inequality on social media emerges through the interaction of platform mechanisms and pre-existing differences in linguistic resources, audience size, institutional support, and content production. Algorithms should not be treated as autonomous agents that intentionally suppress regional languages; nor should unequally search outcomes be reduced to neutral reflections of user preference. Ranking systems process engagement signals, textual metadata, prior circulation, and predicted relevance, all of which are already shaped by unequal sociolinguistic conditions. Indonesian is therefore expected to benefit from a cumulative visibility advantage: broader use generates more searchable content and interactions, which may subsequently strengthen its platform prominence. Regional languages, by contrast, may achieve visibility within narrower thematic domains or through Indonesian subtitles and metadata, producing a form of inclusion that remains dependent on the dominant language. Testing this proposition reframes digital language inequality as a relational process in which technological classification, communicative practice, and national language hierarchy mutually constitute the boundaries of linguistic visibility [10].

2. LITERATURE REVIEW

2.1. Algorithmic visibility

Algorithmic visibility refers to the technologically mediated probability that content, actors, or ideas will become perceptible within platform interfaces. Early accounts conceptualised algorithms as classificatory procedures that prioritise certain information according to encoded criteria, whereas platform scholarship treats visibility as relationally produced through ranking systems, user behaviour, commercial objectives, and content practices [11], [12]. Bucher (2012, 2018) emphasises the power exercised through the threat of invisibility [1], [13], while Cotter (2019) and Bishop (2019) foreground creators’ strategic responses to opaque ranking mechanisms [3], [14]. Visibility is therefore neither a stable technical property nor merely accumulated popularity; it is a contingent outcome of sociotechnical selection.

Algorithmic visibility can be examined through exposure, discoverability, ranking, persistence, and engagement. Exposure concerns whether content enters users’ perceptual fields, while discoverability denotes its retrievability through search, hashtags, or recommendation systems. Ranking differentiates visibility by position, since top-listed content commands greater attention than content appearing deeper within an interface. Persistence captures whether exposure remains stable across sessions, whereas engagement describes interactions that may reinforce subsequent circulation [3], [4]. Research must

consequently distinguish content availability from platform visibility. Counting videos alone cannot establish algorithmic advantage; comparative auditing must examine ranked positions, top-*k* presence, repeated-query stability, and variation across controlled search conditions.

2.2. Linguistic inequality

Linguistic inequality describes the unequal distribution of authority, resources, recognition, and communicative opportunities among languages and their speakers. Classical sociolinguistics associates such inequality with institutional domains, ethnolinguistic vitality, and intergenerational transmission [15], [16]. Critical approaches, however, reject interpretations that attribute language marginalisation solely to declining speaker numbers. They instead examine how political economies assign unequal value to linguistic repertoires. Digital research extends this argument by showing that languages receive highly uneven technological support, data resources, and computational performance [5], [17]. Digital linguistic inequality thus concerns both representational absence and unequal infrastructural accommodation.

Its empirical indicators encompass corpus availability, technological support, domain breadth, institutional representation, audience reach, and conditions of participation. Resource-based approaches classify languages according to datasets, computational tools, and benchmark performance, but risk reducing inequality to technical scarcity [17], [18]. Sociolinguistic approaches instead evaluate whether languages function across education, news, entertainment, commerce, and everyday interaction. Schneider (2022) further demonstrates that commercial technologies regiment multilingualism by privileging standardised, data-rich forms [2]. For platform research, linguistic inequality should therefore be assessed through ranked visibility, thematic diversity, creator ecology, metadata compatibility, and the extent to which regional-language content remains dependent on dominant-language infrastructures.

2.3. Digital linguistic mediation

Digital linguistic mediation refers to the processes through which language becomes reformatted, translated, labelled, and made searchable within technological systems. One perspective regards mediation as expanding access through subtitles, automated translation, transliteration, and multilingual metadata. Another stresses that such mechanisms may reproduce hierarchy when minority-language content must adopt a dominant language to become discoverable [2], [19]. Platformed multilingualism is consequently not equivalent to the simple coexistence of several languages. It is structured by interfaces, classification standards, moderation systems, and monetisation regimes that recognise some linguistic forms more effectively than others. Mediation can increase circulation while simultaneously limiting regional-language autonomy.

Relevant indicators include congruence between spoken language and metadata, subtitle direction, code-switching, translation dependence, and correspondence between linguistic and cultural representation. A video about Javanese culture, for example, does not constitute Javanese-language content when Indonesian dominates its speech and metadata. Conversely, regional-language audio accompanied by Indonesian titles or subtitles represents mediated rather than fully autonomous visibility [20]. Digital-domain breadth must likewise be separated from token representation: frequent appearance within comedy or heritage content may coexist with absence from education, news, or everyday communication. This study therefore theorises platform inequality as the interaction of ranked exposure, functional domain distribution, and dominant-language mediation, rather than as a numerical disparity alone.

3. METHOD

The material units of analysis were publicly accessible YouTube videos provisionally associated with Indonesian, Javanese, or Madurese language use. The corpus comprised 30 candidate videos retrieved on 13 June 2026: ten per target language. Videos were stratified by education, humour, food, local culture, news, and everyday life. Each record contained a unique code, title, URL, search query, retrieval order, snippet evidence, available engagement information, language-verification note, eligibility status, retrieval source, and research notes. Because linguistic identity cannot be inferred reliably from titles or cultural topics alone, spoken-language verification remained a prerequisite for final inclusion.

This research adopted a sequential mixed-method platform-audit design combining controlled algorithmic auditing, stratified quantitative content analysis, and sociolinguistic domain analysis. Algorithm audits are appropriate for examining ranked platform outputs but remain sensitive to account histories, interface configurations, locations, and collection procedures [21]. The stage was therefore used to test queries, refine thematic strata, and establish coding rules. It was not treated as evidence of YouTube amplification because `Result_Order` represented external web-search ordering rather than native platform ranking. Claims concerning algorithmic visibility require a subsequent controlled YouTube search audit.

Table 1. Corpus dataset of research

Corpus code	Target language	Theme	Main characteristics
ID-EDU	Indonesian	Education	Language-learning and school-instruction materials
ID-HUM	Indonesian	Humour	Short comedy and language-based humour
ID-FOD	Indonesian	Food	Culinary education and recipe videos
ID-CUL	Indonesian	Local culture	Cultural education and preservation discourse
ID-DAY	Indonesian	Everyday life	Student and Ramadan daily-life vlogs
JV-EDU	Javanese	Education	School lessons and Javanese-language instruction
JV-HUM	Javanese	Humour	Javanese comedy and regional-language performance
JV-FOD	Javanese	Food	Javanese culinary topics; spoken language unconfirmed
JV-NEWS	Javanese	News	<i>Pawarta</i> and Javanese-language news materials
JV-DAY	Javanese	Everyday life	Everyday-life narrative
JV-CUL	Javanese	Local culture	Cultural representation with textual language evidence
MD-EDU	Madurese	Education	School-based Madurese-language learning
MD-HUM	Madurese	Humour	Comedy with explicit or probable Madurese use
MD-NEWS	Madurese	News	Madurese-language institutional news
MD-CUL	Madurese	Local culture	Cultural events and regional representation

Information was drawn from four interconnected sources within the workbook: the video records, the documented search-query register, the operational codebook, and the dataset summary. Search snippets supplied preliminary evidence concerning language, genre, and topic, while video pages constituted the required source for confirming dominant speech, metadata language, subtitles, publication information, and engagement indicators. The analytical information was organised at video level, but comparisons were conducted across language–theme strata. No creator intention, audience identity, or algorithmic rationale was inferred from metadata. This separation preserved the distinction between observable platform outputs and interpretations of the mechanisms potentially shaping them.

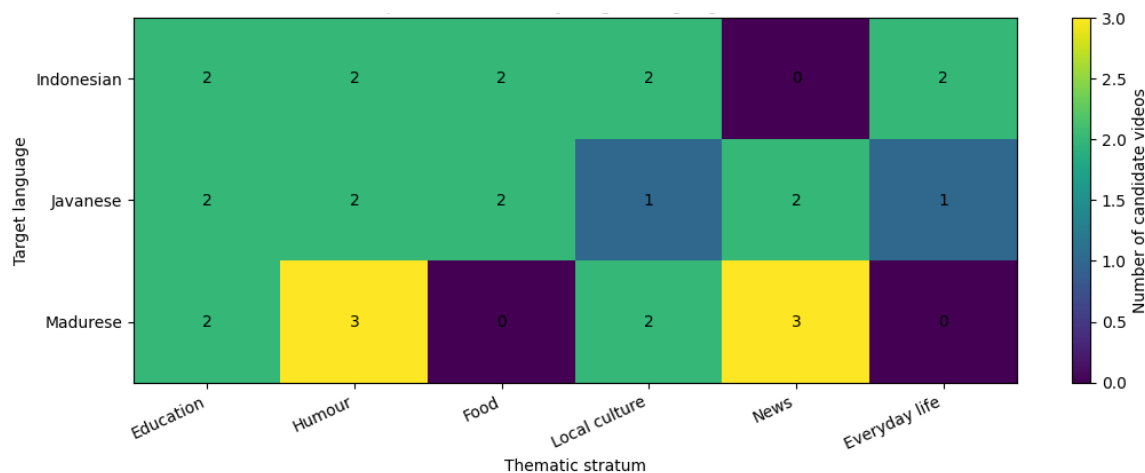
Data collection proceeded through discovery, screening, full coding, and controlled auditing. Discovery employed twelve documented web queries targeting language–theme combinations and produced the 30-candidate corpus. Screening assessed URL accessibility, thematic relevance, duplicate status, and the strength of linguistic evidence. Full coding requires viewing each eligible video to identify dominant spoken language, title and description language, subtitle direction, code-switching, producer type, and communicative domain. The subsequent audit must run semantically comparable queries directly within YouTube under standardised account, location, interface-language, device, and temporal conditions. Ranked outputs and engagement values must be recorded at one documented collection point.

Analysis followed three stages. First, a controlled visibility audit compares top-*k* presence, median position, rank-weighted visibility, and stability across repeated search conditions without interpreting differences as intentional discrimination. Second, stratified content analysis cross-tabulates language with theme, genre, producer type, and communicative function; independently coded variables should be evaluated using Krippendorff's alpha [22]. Third, sociolinguistic analysis examines domain breadth, thematic concentration, spoken–metadata congruence, subtitle direction, and Indonesian-language mediation. UNESCO's new-domain framework informs domain interpretation, but the dataset cannot establish overall language vitality. The three analytical layers jointly distinguish availability, ranked exposure, functional diversity, and mediated linguistic presence.

4. RESULTS

4.1. Retrieval composition and verification readiness

The corpus provides an initial map of how the three target languages were represented across the thematic strata established during the discovery stage. Figure 1 organises the 30 candidate videos by target language, thematic category, retrieval characteristics, and verification status. This structure is analytically necessary because numerical representation alone cannot distinguish broad functional presence from concentration within particular content types. Platform visibility research similarly differentiates the availability of content from its actual discoverability and ranked exposure within platform interfaces [4], [23]. At this stage, Figure 1 therefore serves as a corpus-diagnostic instrument rather than evidence of algorithmic inequality. It identifies the thematic and verification conditions that must be resolved before controlled comparisons of YouTube visibility can be undertaken.



Note: Values represent the 30-video pilot discovery corpus, not native YouTube ranking or platform-wide prevalence.

Figure 1. Corpus distribution by target language and theme

Figure 1 reveals both deliberate balance and internal variation. Indonesian, Javanese, and Madurese each contributed ten candidates, representing 33.33% of the corpus; this equality resulted from the sampling design. Javanese covered all six thematic strata, whereas Indonesian appeared in five and Madurese in four. Indonesian records were evenly distributed across education, humour, food, local culture, and everyday life, with two candidates in each category. Javanese showed two candidates in education, humour, food, and news, but only one in local culture and everyday life. Madurese was concentrated in humour and news, each comprising three records, while food and everyday life were absent. Verification was also uneven: all Indonesian and Javanese records remained provisional, whereas four Madurese records were eligible subject to audio confirmation.

This finding suggests that linguistic presence should be conceptualised through functional breadth and evidential certainty rather than candidate frequency alone. UNESCO's language-vitality framework identifies participation in new domains and media as distinct from mere language existence, implying that concentration in restricted content categories may represent a narrower form of digital functionality. However, the current pattern cannot establish regional-language marginalisation because its thematic structure was shaped by discovery queries and because spoken language remains incompletely verified. Research on multilingual technologies further demonstrates that unequal digital representation emerges through differences in data resources, standardisation, and technological accommodation [2], [5]. This finding therefore establishes an analytical baseline: subsequent auditing must determine whether thematic concentration persists in native YouTube outputs and corresponds with unequal ranked exposure

4.2. Thematic breadth and content concentration across languages

This section examines whether numerical representation in the corpus was accompanied by comparable thematic breadth across Indonesian, Javanese, and Madurese. This distinction is necessary because equal sample sizes may conceal substantial differences in the communicative domains through which languages become digitally represented. Figure 2 therefore disaggregates the ten candidates assigned to each language into six thematic categories and reports both within-language percentages and derived measures of breadth and concentration. Figure 2 reproduces these frequencies without weighting or extrapolation. In platform and minority-language research, domain diversity is analytically distinct from simple content availability because a language may be repeatedly visible while remaining associated with only a restricted set of topics. The analysis consequently evaluates the internal composition of the corpus rather than estimating the overall distribution of YouTube content.

The three language subsets displayed different levels of thematic dispersion. Javanese had the broadest distribution, covering all six categories and producing a thematic breadth score of 100%. Four categories—education, humour, food, and news—contained two records each, while local culture and everyday life contained one each. Indonesian covered five categories, with two records in every represented category and no news candidate. Its distribution was therefore even but not fully comprehensive. Madurese covered four categories and was more concentrated: humour and news each accounted for three candidates, or 30%, whereas education and local culture each accounted for two. No Madurese candidate appeared in food or everyday life. The corresponding concentration indices were 0.18 for Javanese, 0.20 for Indonesian, and 0.26 for Madurese.

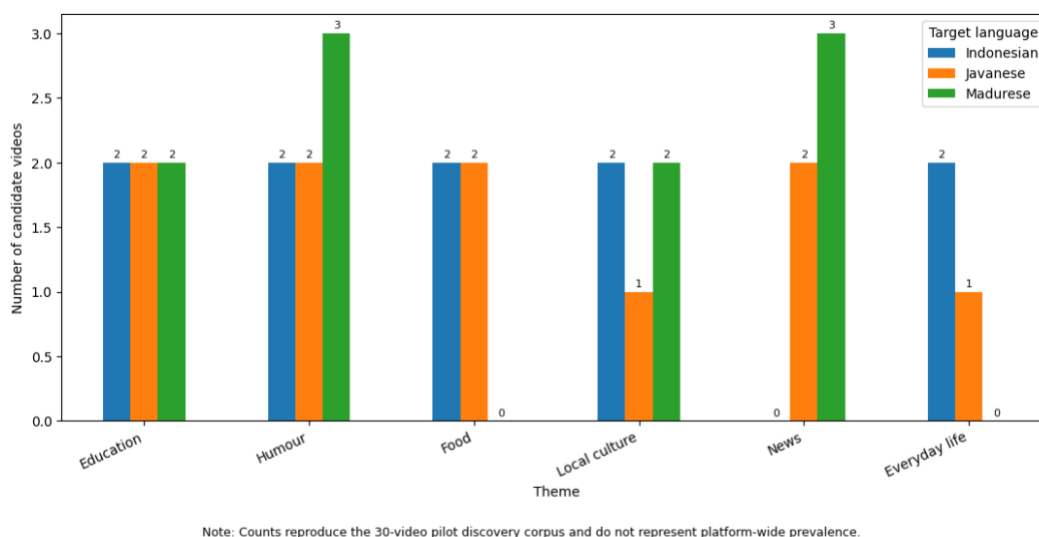


Figure 2. Thematic distribution of the corpus by target language

The variation indicates that digital linguistic presence cannot be reduced to the number of retrieved items; it also concerns the range of social functions in which a language appears. Fishman's domain-based account of language maintenance and UNESCO's framework for language use in new media both suggest that broader functional distribution may provide stronger conditions for sustained public relevance than repeated occurrence within a limited genre. From this perspective, the Javanese subset exhibits the widest - domain coverage, while the Madurese subset appears more strongly associated with humour and institutional news. Nevertheless, this pattern should not be interpreted as intrinsic functional limitation or algorithmic marginalisation. Search-query design, incomplete spoken-language verification, and the small stratified corpus may have shaped the distribution. Result B therefore generates a proposition for subsequent testing: regional-language inequality may involve thematic restriction as well as unequal ranking.

4.3. Indonesian-mediated regional-language visibility and verification

The linguistic classification of platform content requires a distinction between the language targeted by a search query and the language demonstrably used within the retrieved video. Figure 3 therefore evaluates the strength of available linguistic evidence, the readiness of records for inclusion, and documented forms of mediation through Indonesian. This distinction is particularly important in multilingual digital environments, where cultural labels, geographical references, and language names in titles may improve retrievability without corresponding to the dominant spoken language. Research on digitally mediated multilingualism similarly cautions against equating metadata representation with communicative practice, because platform interfaces reformat language through titles, descriptions, captions, translation, and searchable labels [19]. The analysis consequently treats linguistic visibility as an evidentially layered condition rather than a binary classification.

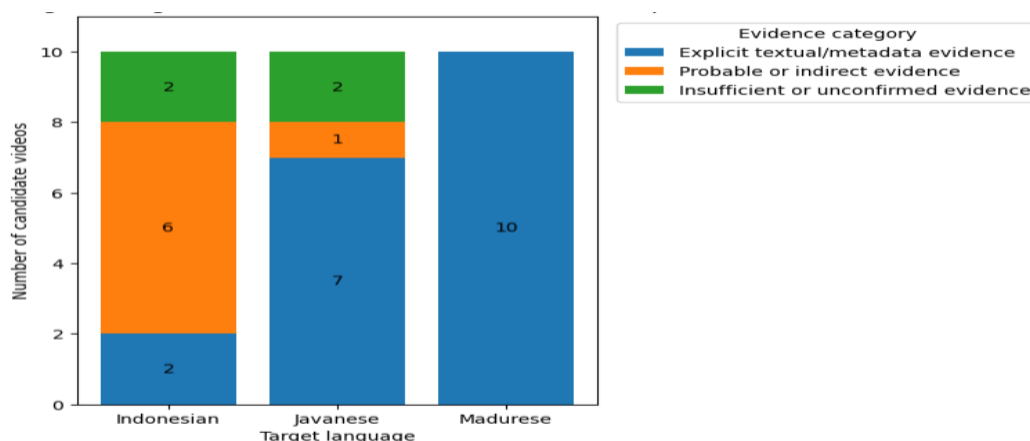


Figure 3. Strength of linguistic verification evidence in the corpus

The strongest documentary evidence occurred in the Madurese subset, where all ten candidates contained an explicit reference to Madurese in their titles, snippets, descriptions, programme labels, or event information. Four were classified as eligible after audio confirmation, while six remained provisional. Javanese contained seven records with explicit evidence, one supported only by search-snippet association, and two food-related videos whose spoken language was unconfirmed. Indonesian displayed the weakest explicit-evidence profile: two records contained direct linguistic references, six were classified as probably Indonesian, and two relied only on titles or snippets. Bilingual mediation was documented in two cases. One Madurese comedy video identified Madurese as its language and Indonesian as its subtitle language, while one Javanese cultural item presented Javanese lyrics with Indonesian translation.

These patterns indicate that regional-language visibility can be simultaneously strengthened and subordinated through dominant-language mediation. Indonesian subtitles or translations may broaden accessibility, yet they can also position Indonesian as the interpretive gateway through which regional-language content becomes intelligible to wider audiences. This corresponds with accounts of platformed multilingualism that view translation and metadata not as neutral additions but as classificatory infrastructures that distribute linguistic recognition unevenly [2], [24]. The pattern also exposes a methodological risk: culturally Javanese or Madurese content may be counted as regional-language material even when Indonesian dominates its speech. Accordingly, the relevant inequality is not simply presence versus absence. It concerns whether a language is independently audible, searchable, and functionally usable, or visible mainly through labels, subtitles, and metadata supplied in the nationally dominant language.

5. DISCUSSION

The distribution demonstrates that equal numerical representation does not necessarily produce equivalent linguistic visibility. Each language occupied one-third of the corpus, yet the evidence supporting its inclusion varied substantially, revealing a functional distinction between sampled presence and verified platform presence. This matters because research on algorithmic visibility has repeatedly shown that content availability is not synonymous with discoverability, prominence, or sustained exposure [1], [3], [4]. A balanced corpus may therefore conceal asymmetries in how confidently content can be classified, how consistently it appears across domains, and how effectively it can be retrieved through platform metadata. The practical implication is methodological as much as substantive: studies of linguistic inequality risk reproducing platform classifications when they treat search labels as transparent evidence of language use. The present findings thus reposition visibility as a layered construct involving retrieval, verification, ranking, and communicative function rather than a simple count of videos associated with each language.

The uneven verification profile appears to reflect differences in how languages are named, labelled, and institutionalised within searchable content. Indonesian was frequently inferred from communicative context rather than explicitly marked because it operates as the national default and therefore often remains unlabelled. Regional languages, by contrast, were more likely to be named in titles, programme descriptions, or educational labels precisely because their use is marked and must be specified. This interpretation resonates with Schneider's (2022) argument that digital systems regiment multilingualism through standardised labels and data practices, but it also complicates resource-centred accounts of technological inequality [2]. Explicit regional-language metadata may improve retrievability even when broader technological support remains weak. Conversely, the apparent invisibility of Indonesian as a named language may coexist with its structural dominance as the unmarked communicative norm. The pattern therefore suggests correlation between linguistic markedness and metadata explicitness, not direct algorithmic preference. Native platform auditing is required to determine whether such metadata differences translate into unequal ranked exposure.

The thematic distribution further indicates that linguistic inequality may operate through functional differentiation rather than outright exclusion. Javanese appeared across all six categories, Indonesian across five, and Madurese across four, with the latter more concentrated in humour and news. Such variation matters because domain breadth contributes to whether a language is encountered as a flexible medium for education, information, culture, entertainment, and everyday interaction. Fishman's domain perspective and UNESCO's framework for new media both imply that repeated visibility within a limited set of functions differs from participation across multiple social domains [15]. The pattern therefore raises the possibility that regional languages may remain publicly audible while being channelled into recognisable cultural or performative niches [25]. This interpretation is consistent with minority-language media research that associates vitality with functional expansion, yet it also cautions against reading thematic concentration as linguistic weakness. Concentration can reflect institutional investment, creator specialisation, or search design rather than reduced expressive capacity.

The underlying structure of thematic variation is likely relational rather than intrinsic to the languages themselves. Content production depends on institutional sponsorship, audience expectations, genre

conventions, and the availability of searchable terminology. Madurese news candidates, for example, were associated with explicitly labelled regional-language programming, suggesting that institutional broadcasting can create concentrated but recognisable visibility. Javanese content, by contrast, extended across educational, humorous, culinary, journalistic, cultural, and everyday genres, potentially reflecting a larger production ecology and more diversified creator base. This interpretation aligns with Blasi, Anastasopoulos, and Neubig's (2022) demonstration that technological inequality follows uneven resource distribution [5], but the present findings suggest that resource differences are also organised by genre and institution. At the same time, the absence of Madurese food or everyday-life candidates cannot be treated as platform suppression because the queries and small sample shaped retrieval. The pattern identifies a structural hypothesis: domain inequality may emerge from unequal production infrastructures before algorithmic ranking intensifies it.

The mediation evidence shows that regional-language visibility may depend on Indonesian as an access language. Indonesian subtitles accompanying Madurese speech and Indonesian translations accompanying Javanese lyrics can enlarge potential audiences, facilitate comprehension, and improve circulation beyond local-language communities. In this sense, mediation performs an inclusive function by connecting regional-language content to the national public sphere. Yet it can also produce a hierarchical arrangement in which regional languages remain audible while Indonesian controls explanation, searchability, and interpretive accessibility. Kelly-Holmes (2019) and Schneider (2022) similarly argue that digital multilingualism is structured through technological and commercial infrastructures rather than constituted by simple language coexistence [2], [19]. The implication is that inclusion should not be assessed solely by whether regional languages appear on screen or in audio. A more consequential question is whether they can function autonomously across speech, metadata, captions, and discovery. The findings therefore distinguish linguistic presence from infrastructural independence without assuming that translation itself is inherently marginalising.

This mediation pattern is plausibly linked to the asymmetrical sociolinguistic position of Indonesian and regional languages. Indonesian offers the widest national reach, stronger institutional recognition, and greater compatibility with dominant platform conventions, making it a practical choice for subtitles, titles, and descriptions. Regional-language creators may therefore adopt Indonesian metadata strategically to increase discoverability rather than because platforms explicitly require it. Cotter's (2019) account of visibility games is relevant here: creators adapt content presentation to anticipated algorithmic and audience expectations under conditions of opacity [3]. However, strategic adaptation can cumulatively reinforce the dominant language by generating more searchable Indonesian text around regional-language content. The resulting inequality is recursive: Indonesian facilitates access to regional-language videos while simultaneously accumulating further metadata, engagement, and platform legibility. The evidence cannot establish that YouTube causally produces this cycle, but it supports a relational model in which creator strategies, sociolinguistic hierarchy, and platform classification interact to shape unequal forms of linguistic visibility.

6. CONCLUSION

This study shows that linguistic inequality on YouTube cannot be understood through numerical presence alone. The central lesson is that visibility is multidimensional: it depends on retrieval, thematic breadth, evidential certainty, metadata congruence, and the extent to which regional-language content relies on Indonesian for subtitles, description, and discoverability. Its principal contribution lies in integrating algorithmic visibility auditing, stratified content analysis, and sociolinguistic mediation analysis within one framework. This approach extends existing platform research by treating language inequality not simply as absence, but as uneven ranked exposure, restricted functional distribution, and dependence on dominant-language infrastructure.

The study remains limited by its small corpus, externally retrieved search order, incomplete audio verification, uneven engagement metadata, and absence of repeated native YouTube audits. These constraints prevent platform-wide generalisation and causal claims about algorithmic suppression. Future research should therefore employ controlled YouTube searches across multiple sessions, devices, locations, and account conditions, while expanding the corpus and verifying spoken language, metadata language, subtitles, producer type, and engagement trajectories. Comparative longitudinal studies involving additional Indonesian regional languages would also clarify whether the observed patterns reflect stable platform structures, temporary ranking conditions, or differences in production ecosystems.

ACKNOWLEDGMENTS

The author gratefully acknowledges colleagues and academic peers who provided valuable feedback and support during the preparation of this study.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Daroji: conceptualization (lead), methodology (lead), formal analysis (lead), writing – original draft (lead), writing – review and editing (lead).

CONFLICT OF INTEREST STATEMENT

Author state 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

I have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

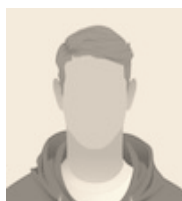
Data availability is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

- [1] T. Bucher, *If... then: Algorithmic power and politics*. Oxford University Press, 2018.
- [2] B. Schneider, "Multilingualism and AI: The regimentation of language in the age of digital capitalism," *Signs and Society*, vol. 10, no. 3, pp. 362–387, 2022, doi: 10.1086/721757.
- [3] K. Cotter, "Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram," *New Media & Society*, vol. 21, no. 4, pp. 895–913, 2019, doi: 10.1177/1461444818815684.
- [4] B. E. Duffy and C. Meisner, "Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility," *Media, Culture & Society*, vol. 45, no. 2, pp. 285–304, 2023, doi: 10.1177/01634437221111923.
- [5] D. E. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic inequalities in language technology performance across the world's languages," presented at the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, 2022, pp. 5486–5505. doi: 10.18653/v1/2022.acl-long.376.
- [6] L. Song, X. Hu, G. Zhang, P. Spachos, K. N. Plataniotis, and H. Wu, "Networking systems of AI: On the convergence of computing and communications," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20352–20381, 2022.
- [7] A. Fawaid, *Who Controls the Narrative? Language, Media, and Politics of Representation*. Probolinggo: Pustaka Nurja, 2026.
- [8] N. H. Mazlan, C. W. Putra, and H. Sulisty, "Understanding reader navigation patterns in multi-path hypertext fiction: A case study approach to Patchwork Girl," *lingtech*, vol. 1, no. 2, pp. 1–12, Jun. 2025, doi: 10.64595/lingtech.v1i1.3.
- [9] M. R. M. Nor and A. Zubaidi, "Digital reading platforms in literary learning: cultural understanding and engagement in Indonesian and Malay texts," *Lingua Technica: Journal of Digital Literary Studies*, vol. 2, no. 1, pp. 33–50, 2026, doi: 10.64595/lingtech.v2i1.131.
- [10] J. Leung, "Hurdles and horizons of linguistics for social justice," *Language in Society*, vol. 52, pp. 882–893, Nov. 2023, doi: 10.1017/S0047404523000635.
- [11] T. Gillespie, "The relevance of algorithms," in *Media technologies: Essays on communication, materiality, and society*, T. Gillespie, P. J. Boczkowski, and K. A. Foot, Eds., MIT Press, 2014, pp. 167–194.
- [12] R. Kitchin, "Thinking critically about and researching algorithms," *Information, Communication & Society*, vol. 20, no. 1, pp. 14–29, 2017, doi: 10.1080/1369118X.2016.1154087.
- [13] T. Bucher, "Want to be on the top? Algorithmic power and the threat of invisibility on Facebook," *New Media & Society*, vol. 14, no. 7, pp. 1164–1180, 2012, doi: 10.1177/1461444812440159.
- [14] S. Bishop, "Managing visibility on YouTube through algorithmic gossip," *New Media & Society*, vol. 21, no. 11–12, pp. 2589–2606, 2019, doi: 10.1177/1461444819854731.
- [15] J. A. Fishman, *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters, 1991.
- [16] UNESCO Ad Hoc Expert Group on Endangered Languages, "Language vitality and endangerment," UNESCO, 2003.
- [17] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 6282–6293. doi: 10.18653/v1/2020.acl-main.560.
- [18] A. Grütznher-Zahn, "Surveying the technology support of languages," presented at the Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages, Association for Computational Linguistics, 2024.
- [19] H. Kelly-Holmes, "Multilingualism and technology: A review of developments in digital communication from monolingualism to idiolingualism," *Annual Review of Applied Linguistics*, vol. 39, pp. 24–39, 2019, doi: 10.1017/S0267190519000102.

- [20] R. W. Ikomah and Z. H. Sain, "Text mining and semantic modeling of literary corpora: a machine learning-based study of Indonesian fiction," *Lingua Technica: Journal of Digital Literary Studies*, vol. 2, no. 1, pp. 51–67, 2026, doi: 10.64595/lingtech.v2i1.133.
- [21] E. Harmer and R. Southern, "Twitter to Yell at Them, Facebook to Ask Questions and Instagram to Show Support? Exploring Platform Effects in Responses to Women Politicians On Social Media," *Political Communication*, pp. 1–21, Nov. 2025, doi: 10.1080/10584609.2025.2585498.
- [22] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004, doi: 10.1111/j.1468-2958.2004.tb00738.x.
- [23] 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.
- [24] 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.
- [25] R. Assyabani, "Multimodal poetics in digital literature: A corpus-based analysis of visual-verbal design, screen-based textuality, and reader interaction," *Lingua Technica: Journal of Digital Literary Studies*, vol. 1, no. 1, pp. 41–53, Jan. 2025, doi: 10.64595/cy0q2n14.

BIOGRAPHY OF AUTHOR



Daroji    is affiliated with Universitas Sarjanawiyata Tamansiswa, Indonesia. His research focus are on language, culture, and related humanities disciplines. His scholarly research include discourse studies, cultural representation, spatial language, architectural narratives, and the relationship between language and heritage. His work often examines how traditional knowledge and cultural identity are articulated through language and descriptive discourse. He can be contacted at email: daroji250@gmail.com