Indonesian Journal of Computational Language Studies
Vol. 1, No. 3, September 2026, pp. 164~177
ISSN: 3163-7884, DOI: https://doi.org/10.67490/ijcl.v1i3.1027

Code-switching as a computational problem: modeling Indonesian–English digital talk beyond monolingual assumptions

Devita Septian Dwi Hidayati
Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia

Article Info

Article history:

Received Jul 17, 2026
Revised Aug 12, 2026
Accepted Sep 25, 2026

Keywords:

code-switching;
computational sociolinguistics;
Indonesian–English;
language identification;
multilingual NLP.

ABSTRACT

Background: Indonesian–English digital talk increasingly exposes the limits of language technologies that assume stable monolingual input, because switch-points often carry stance, affect, technical meaning, platform practice, and identity across compact social media utterances. **Objective:** This study examines how Indonesian–English code-switching can be modeled as a computational problem without reducing mixed-language digital utterances to noise, error, or preprocessing residue. **Method:** Using a corpus-based qualitative-computational design, this study analyzes 15 coded public digital text records through token-level language identification, switch-point mapping, normalization–translation alignment, and pragmatic-computational risk coding. **Results:** Code-switching appears as structured multilingual practice, with English segments functioning as evaluative phrases, technical labels, platform commands, fandom markers, and discourse-sequence devices within Indonesian frames. Normalization and translation are shown to be interpretive operations, because full correction or translation may flatten sentiment scope, register, and locally meaningful hybrid forms. Monolingual processing risks forced single-language labeling, technical mistranslation, discourse-sequence loss, and representational distortion for multilingual users. **Implication:** These findings reposition computational adequacy as a question of meaning preservation, not merely model accuracy or benchmark coverage in a low-resource Indonesian digital context for language technologies. **Novelty:** This study offers an integrated framework that links linguistic switching structure, preprocessing risk, and accountable multilingual NLP representation.

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

Corresponding Author:

Devita Septian Dwi Hidayati
Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia
Jl. Raya Rungkut Madya, Gunung Anyar, Surabaya, 60294, Jawa Timur, Indonesia
Email: 24046010029@student.upnjatim.ac.id

1. INTRODUCTION

Indonesia's digital public sphere has become a dense multilingual environment in which Indonesian, English, regional languages, platform slang, emojis, abbreviations, and algorithmic templates circulate within the same communicative space. This condition is not marginal: APJII reported 221,563,479 internet users in Indonesia in 2024, equal to 79.5% of population, while DataReportal [1] recorded 143 million social media user identities in Indonesia in January 2025. These figures make Indonesian digital talk an empirically consequential site for studying how language technologies process everyday bilingual expression. Yet computational systems often still presume that input belongs to one stable language, even when users routinely cross lexical, syntactic, and pragmatic boundaries. Indonesian–English code-switching is therefore crucial not because it is linguistically unusual, but because it exposes a mismatch between lived multilingual

communication and monolingual NLP design. This mismatch becomes visible in tokenization, language identification, normalization, sentiment analysis, and output alignment.

Existing scholarship has established important foundations for computational Indonesian language processing. Santoso [2] addresses named entity recognition through BiLSTM-CRF and domain-specific embeddings, Novitasari [3] examines transformer-based sentiment analysis in Indonesian political discourse, Wahid [4] develops corpus-based lexical analysis for Indonesian academic writing, and Yuliyana [5] evaluates readability through hybrid NLP approaches. More directly, Nugroho [6] investigates Indonesian–English code-mixing detection through machine learning and linguistic features, while Barik et al.'s [7] Indonesian–English Twitter work introduced a pipeline involving tokenization, language identification, lexical normalization, and translation. Recent multilingual studies also show growing concern with low-resource Indonesian dialects, culturally grounded LLM evaluation, multilingual reasoning, and sociolinguistic bias [8], [9], [10], [11]. However, what remains underdeveloped is an integrated account of code-switching as both a sociolinguistic practice and a computational modeling problem. Much prior work detects mixed language; less explains what monolingual assumptions erase.

This study addresses that gap by asking how Indonesian–English digital code-switching can be modeled without reducing multilingual talk to noise, error, or preprocessing residue. Specifically, this study examines three linked questions: how switch-points are structured across Indonesian–English digital utterances; how normalization and translation procedures preserve or flatten pragmatic meaning; and what computational risks emerge when mixed-language data are processed through monolingual or language-dominant assumptions. These questions position code-switching not merely as an object of linguistic classification but as a stress test for NLP systems designed around stable language boundaries. This study therefore connects token-level language identification with discourse-pragmatic interpretation, asking whether computational models can recognize not only which language appears, but also what communicative work switching performs. In methodological terms, this study combines corpus metadata, text-level documentation, and coding categories for Result A, Result B, and Result C, aligning linguistic description with computational evaluation rather than treating them as separate tasks.

This study advances a provisional argument that Indonesian–English code-switching should be modeled as structured multilingual practice rather than as deviation from standardized Indonesian or English. This argument follows recent critiques of multilingual AI evaluation showing that model performance cannot be understood only through benchmark accuracy, because cultural grounding, local-language safety, hallucination, and representational bias also shape system reliability [9], [12], [13], [11]. It also aligns with work on parameter-efficient adaptation for low-resource Indonesian languages, which suggests that language technologies must be adapted to sociolinguistic diversity rather than assuming that multilingual pretraining automatically solves local complexity [14]. The expected implication is conceptual and methodological: Indonesian–English digital talk requires models that preserve switch-points, ambiguity, pragmatic function, and user identity. Thus, this study tests whether computational treatment of code-switching can move beyond language detection toward accountable multilingual representation.

2. LITERATURE REVIEW

2.1. Code-switching

Code-switching is commonly understood as alternation between two or more languages within a communicative event, yet recent scholarship treats it less as a mechanical shift between bounded codes than as a situated practice shaped by identity, audience, medium, and interactional purpose. Sociolinguistic studies distinguish code-switching from code-mixing, borrowing, and translanguaging, although these boundaries often blur in digital Indonesian–English discourse. Doğruöz et al. [15] and Rayo et al. [16] show that language alternation involves linguistic structure and social meaning, while Azzahra [17], Hanafiah [18], and Sahib et al. [19] demonstrate that Indonesian–English switching operates across public communication, elite interaction, and syntactic configuration. Thus, code-switching is not merely bilingual substitution but patterned multilingual action.

Its main dimensions can be organized around structure, function, and medium. Structurally, studies identify inter-sentential switching, intra-sentential switching, tag switching, lexical insertion, hybrid morphology, and ambiguous mixed forms. Functionally, Indonesian research links switching to emphasis, intimacy, pedagogical mediation, professional identity, humor, prestige, and communicative efficiency [20], [21], [22], [23]. Digitally, platform affordances add hashtags, emojis, abbreviations, usernames, and algorithmic templates to conventional language alternation [24], [25], [26]. These indicators suggest that Indonesian–English code-switching requires analysis beyond sentence-level classification.

2.2. Computational modeling

Computational modeling refers to procedures through which multilingual textual behavior is transformed into machine-readable representations, labels, predictions, or generated outputs. In code-switching research, however, modeling varies from narrow detection to richer representational learning. Nugroho [6] foregrounds machine learning and linguistic features for Indonesian–English code-mixing detection, while Hidayatullah et al. [27] emphasize corpus creation and language identification for Indonesian–Javanese–English tweets. Broader work questions whether multilingual models actually process switching effectively rather than merely absorbing multilingual vocabulary [28], [29]. This distinction is critical: detecting mixed tokens is not equivalent to modeling multilingual meaning.

Computationally relevant indicators include tokenization stability, token-level language identification, switch-point recognition, lexical normalization, named entity preservation, sentiment drift, translation alignment, and robustness under noisy input. Santoso [2], Novitasari [3], Wahid [4], and Yuliyana [5] represent important progress in Indonesian NLP, but their tasks remain vulnerable when digital input contains mixed-language structures. Hidayatullah et al. [30] extend transformer modeling to Indonesian, Javanese, and English code-mixed text, while Rathnayake et al. [31], Mondal et al. [32], and Xie et al. [33] show that multilingual adaptation, controllable generation, and large-scale datasets are increasingly central. The unresolved issue is whether these models preserve pragmatic switching functions.

2.3. Monolingual assumptions

Monolingual assumptions describe a design logic in which language data are treated as belonging primarily to one stable linguistic system. Such assumptions persist in language identification, preprocessing, benchmark construction, sentiment classification, readability assessment, and LLM evaluation. For Indonesian digital communication, this logic is insufficient because English may appear as stance marker, technical register, affective intensifier, quotation, or identity cue rather than as a separate document language. Recent LLM studies sharpen this critique: Mohamed et al. [34], Sheth et al. [35], Winata et al. [36], and Oh et al. [37] suggest that code-switched understanding, reasoning, generation, and output alignment remain unresolved challenges. Monolingual processing therefore risks flattening communicative meaning.

Indicators of monolingual bias include forced single-language labels, misread named entities, mistranslated hybrid expressions, sentiment reversal, loss of local cultural reference, and unequal performance for multilingual users. Indonesian AI scholarship extends this concern from technical accuracy to representation and accountability. Bouazzaoui [9] critiques translated benchmarks, Wulandari [10] examines multilingual reasoning across Indonesian languages, Nadiyah [12] addresses safety failure in local languages, Wibisono [13] analyzes cultural hallucination, Hadziqi [14] explores parameter-efficient adaptation, and Sumadi [11] foregrounds sociolinguistic bias. This study therefore positions Indonesian–English digital code-switching as a test case for accountable multilingual NLP, where linguistic structure, pragmatic function, and computational reliability must be modeled together.

3. METHOD

The unit of analysis in this study is a text record from Indonesian–English digital communication, treated as a materially bounded utterance rather than as evidence about individual users. Primary records are drawn from a public Indonesian–English code-mixed tweet corpus, while secondary and contextual materials document annotation logic, multilingual NLP constraints, and Indonesian AI evaluation debates. This focus is appropriate because code-switching in digital text becomes computationally observable through tokens, language labels, normalization choices, translation equivalents, and discourse context. Prior work on Indonesian code-mixed corpora, multilingual transfer, and code-switching technologies supports text-level and token-level analysis as analytically defensible units [27], [6], [28].

This study uses a corpus-based qualitative-computational design. It does not test causal effects, infer user motivation, or generalize statistically to all Indonesian social media users. Instead, it examines how Indonesian–English switching becomes structured as data when processed through metadata registration, text extraction, token-level annotation, normalization, translation, and computational-risk coding. This design combines linguistic description with methodological auditing: linguistic analysis identifies switching patterns and pragmatic functions, whereas computational analysis evaluates where monolingual preprocessing may distort mixed-language meaning. Such a design follows recent code-switching scholarship that treats multilingual digital talk as structured interaction and recent NLP work that questions whether multilingual systems adequately represent code-switched input.

Table 1. Dataset composition

Code	Data Type	Source	Location / Scope	Period	Quantity	Characteristics	Status
IEDT-P	Primary text records	Indonesian-English Code-Mixed Tweet Dataset	Indonesia-centered digital talk	Source date not provided; retrieved 2026	15 coded records	Public text rows containing Indonesian–English alternation, token labels, normalization, and translation fields	Included
IEDT-S	Secondary methodological sources	ACL / computational code-switching literature	International and Indonesian NLP scholarship	2019–2026	3 sources	Documents corpus design, code-mixing detection, multilingual modeling, and robustness framing	Included
IEDT-C	Contextual dataset and AI evaluation sources	Zenodo, Hugging Face, arXiv, user-provided bibliography	Indonesian multilingual and low-resource NLP context	2022–2026	2 sources	Supports comparison with Indonesian–Javanese–English labeling, code-mixed robustness, and sociolinguistic bias debates	Included

Information sources were organized into three evidentiary layers. Primary data consist of public tweet records from an Indonesian–English code-mixed corpus. Secondary sources include methodological and computational studies on tokenization, language identification, normalization, translation, code-mixing detection, transformer-based modeling, and multilingual language-model evaluation. Contextual sources include Indonesian multilingual datasets and recent Indonesian AI studies on low-resource language modeling, culturally grounded evaluation, hallucination, safety, and sociolinguistic bias. This layered design prevents conflating corpus evidence with interpretive literature: primary records supply analysable text, secondary sources justify analytical procedures, and contextual sources clarify broader computational risks in Indonesian multilingual NLP.

Data collection proceeded through source registration, eligibility screening, metadata extraction, text documentation, and deduplication. Each item received a unique document code, source role, title, retrieval date, source URL, inclusion basis, exclusion status, and deduplication key. Text records were stored with raw text, selected excerpt, token sequence, language-label sequence, normalization note, translation note, and contextual description. Visual and video fields were retained in corpus structure but marked not applicable because this dataset is text-based. Duplicate control was conducted by comparing document code, source URL, and text-based deduplication keys. This procedure preserves auditability while avoiding unsupported reconstruction of missing platform metadata.

Data analysis was conducted in three connected stages. First, token-level language identification and switch-point mapping identified Indonesian, English, ambiguous, named-entity, punctuation, and other tokens, then located switching across lexical, phrasal, and clausal boundaries. Second, normalization and translation alignment audited whether informal spellings, abbreviations, product terms, named entities, affective markers, and English evaluative phrases were preserved or flattened. Third, pragmatic-computational coding connected linguistic function to modeling risk, including language-label ambiguity, sentiment-scope error, entity confusion, mistranslation, and loss of discourse stance. Figure 2 presents this operational matrix as an analytic sequence rather than a causal model.

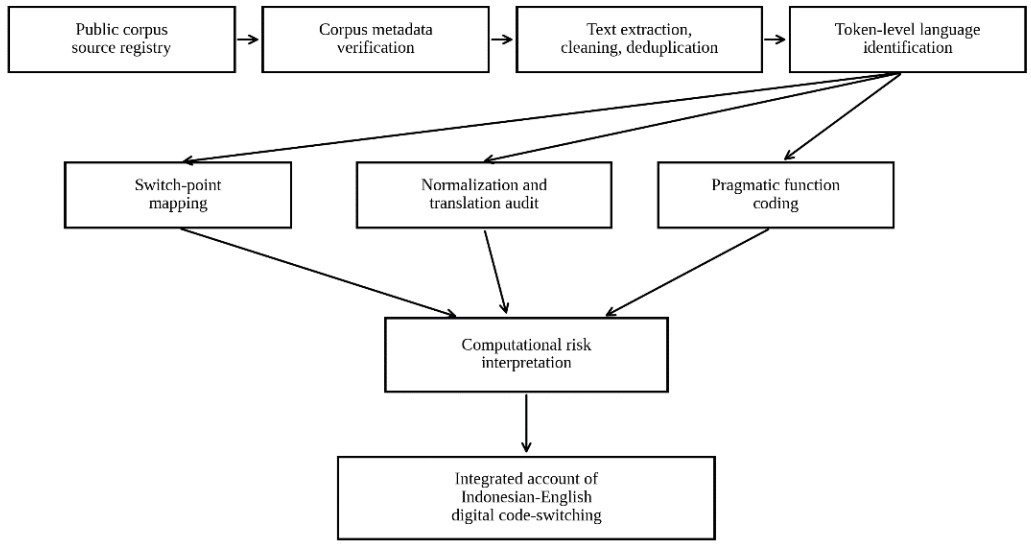

Figure 1. Operational matrix of analysis

4. RESULTS

4.1. Switch-point structure and language-label ambiguity

Indonesian–English digital switching is distributed across multiple configurations rather than concentrated in a single bilingual pattern. The table shows that intra-sentential lexical insertion, evaluative phrase insertion, and inter-sentential switching each appear in two records, while other configurations occur once. This distribution matters because each configuration creates a different computational problem: lexical insertions test token-level language identification, inter-sentential switching tests clause segmentation, and platform or technical insertions test domain-sensitive normalization. The coded records also show that English tokens are not merely foreign insertions; they carry stance, product meaning, fandom identity, platform action, or technical specificity. This pattern aligns with computational code-switching research that treats mixed-language input as structured linguistic data rather than accidental noise.

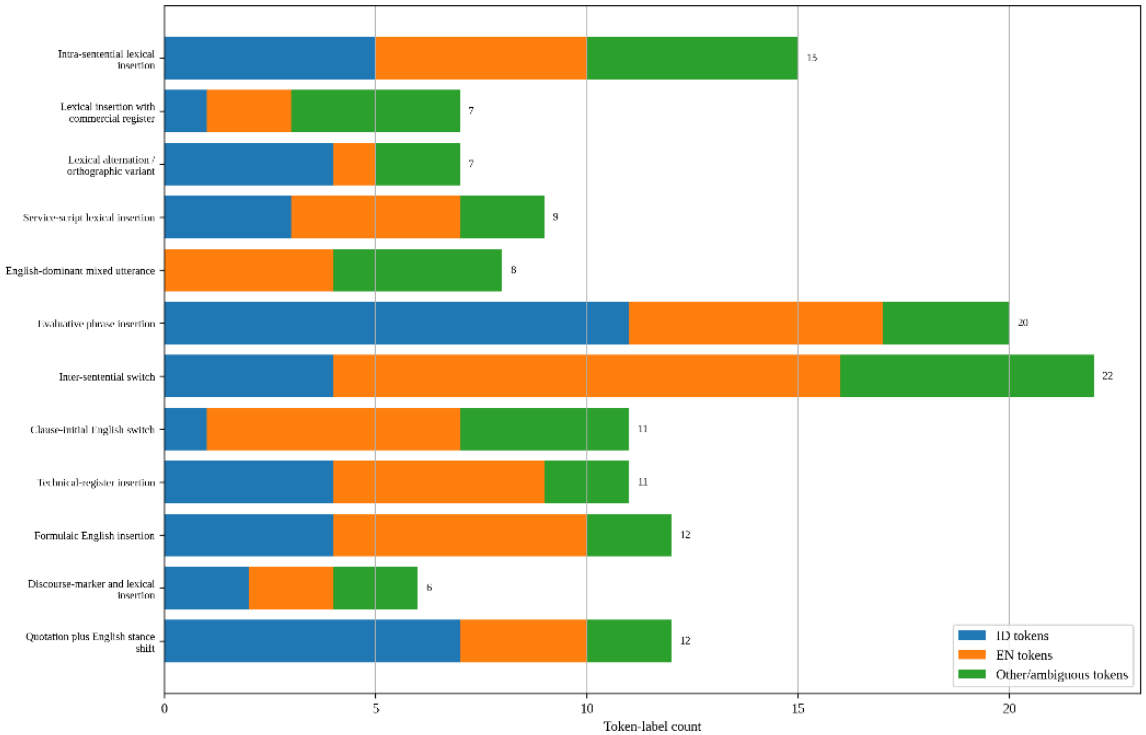

Figure 2. Token-label composition across Indonesian–English switching configurations

Table 2. Switching configurations, token-label composition, and pragmatic orientation

Switching Configuration	Operational Indicator	Frequency	Percentage	ID Tokens	EN Tokens	Other / Ambiguous Tokens	Dominant Pragmatic Orientation	Representative Excerpt	Data Code
Intra-sentential lexical insertion	English phrase embedded beside Indonesian lexical frame; Indonesian matrix with English fandom terms	2	13.3	5	5	5	Compressed evaluation; fandom affiliation	"*not habits.. tekanan*"; "*first luv ku di dunia fangirl*"	IEDT-P-001; IEDT-P-002
Lexical insertion with commercial register	English product terms inserted into Indonesian sales frame	1	6.7	1	2	4	Selling; cataloguing	"*handcream ... photo set ... ada shopee jg*"	IEDT-P-003
Lexical alternation / orthographic variant	Indonesian question frame with Indonesianized and English-like lexical variants	1	6.7	4	1	2	Metalinguistic inquiry	"*Apa bedanya donut sama doughtnut?*"	IEDT-P-004
Service-script lexical insertion	Indonesian institutional reply with English platform actions	1	6.7	3	4	2	Customer assistance; directive	"*respon via DM ... follow back*"	IEDT-P-005
English-dominant mixed utterance	English frame anchored by Indonesia-related named entities	1	6.7	0	4	4	Review; stance-taking	"Penang Food Rocks! Even in Jakarta!"	IEDT-P-006
Evaluative phrase insertion	Indonesian matrix with English sentiment or assessment phrase	2	13.3	11	6	3	Stance; affect; self-assessment	"*enak not bad lah tapi haram*"; "*not so good juga sih*"	IEDT-P-007; IEDT-P-012
Inter-sentential switch	Indonesian narrative followed by English evaluation; English affective opening followed by Indonesian continuation	2	13.3	4	12	6	Judgment; stance intensification; emotion narration	"it's not a good example definitely"; "Its kinda weird while you chat"	IEDT-P-008; IEDT-P-015
Clause-initial English switch	English correction followed by Indonesian explanation	1	6.7	1	6	4	Disagreement; correction	"*No... not like that... kalau Anda sering hunting*"	IEDT-P-009
Technical-register insertion	Indonesian complaint frame with English technical terms	1	6.7	4	5	2	Complaint; troubleshooting	"*speedtest ... server not found ... inbound*"	IEDT-P-010
Formulaic English insertion	Indonesian evaluation with English media-sharing template	1	6.7	4	6	2	Recommendation; affective sharing	"*trending di spotify ... Heres a song for you*"	IEDT-P-011
Discourse-marker and lexical insertion	Indonesian argument with English marker and content noun	1	6.7	2	2	2	Advice; moral evaluation	"*Well tergantung ... skill apa2*"	IEDT-P-013
Quotation plus English stance shift	Indonesian quoted trigger followed by English rebuttal	1	6.7	7	3	2	Irony; rebuttal	"Well at least some generation can take jokes"	IEDT-P-014

The dominant observable pattern is the uneven interaction between language labels and communicative function. English-heavy configurations are clearest in inter-sentential switching, where English tokens reach 12 across two records, and in formulaic or clause-initial switching, where English phrases organize recommendation or disagreement. In contrast, evaluative phrase insertion contains more Indonesian tokens, yet its pragmatic force depends on English expressions such as "not bad" and "not so good." The excerpt "*enak not bad lah tapi haram*" is especially diagnostic because Indonesian supplies moral-affective framing while English supplies polarity. Similarly, "server not found" cannot be translated as ordinary English without losing its technical function. These patterns show why document-level language labels would underdescribe the corpus.

Code-switching here operates as a site where computational representation and sociolinguistic meaning intersect. The recurring presence of English in stance, technical, platform, and affective positions suggests that switch-points are not random disruptions in Indonesian text. They mark discourse functions that may be lost if preprocessing normalizes the record into one language. This supports recent critiques of multilingual modeling which argue that multilingual pretraining and code-switching comprehension are not equivalent capacities, especially when pragmatic meaning depends on mixed-language sequencing [28], [34], [35], [36]. This study therefore treats switch-point structure as a representational problem: models must preserve language alternation, ambiguity, and discourse function together.

4.2. Normalization, translation alignment, and meaning preservation

Indonesian–English digital records reveal that normalization and translation are not neutral preprocessing stages. Each alignment domain shows a different point at which computational handling may alter meaning: affective evaluation depends on polarity and stance, technical/platform terms depend on domain specificity, and commercial or fandom expressions depend on register retention. The highest-frequency domains each contain three records, showing that meaning preservation is distributed across affect, platform practice, and identity-oriented lexical choice rather than confined to one type of switch. This pattern is consistent with computational code-switching work that treats language identification, normalization, and translation as connected tasks, not isolated modules [27], [6], [28].

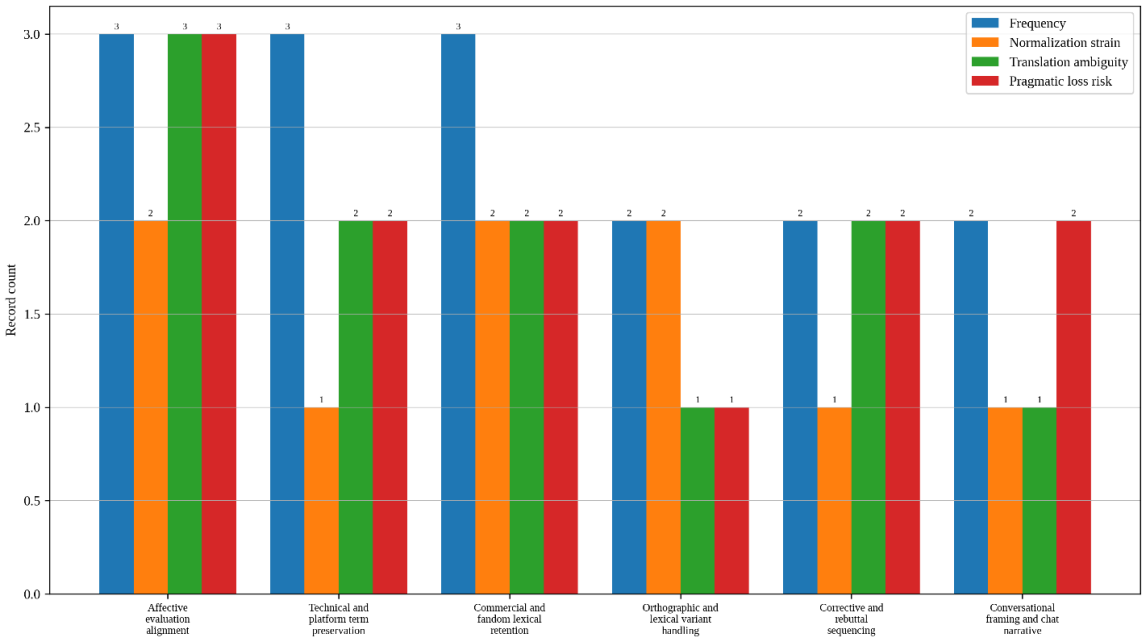

Figure 3. Normalization and translation alignment risks across corpus domains

Table 3. Normalization and translation alignment across Indonesian–English digital records

Alignment Domain	Operational Indicator	Frequency	Percentage	Normalization Strain	Translation Ambiguity	Pragmatic Loss Risk	Dominant Computational Issue	Representative Excerpt	Data Code
Affective evaluation alignment	English evaluative phrase carries sentiment inside Indonesian moral or experiential frame	3	20.0	2	3	3	Polarity may be preserved lexically but misread pragmatically	"*enak not bad lah tapi haram*"; "*not so good juga sih*"; "not a good example definitely"	IEDT-P-007; IEDT-P-012; IEDT-P-008
Technical and platform term preservation	English technical or platform terms function as task-specific commands or diagnostic labels	3	20.0	1	2	2	Literal translation may flatten platform or troubleshooting function	"*respon via DM*"; "server not found"; "Heres a song for you"	IEDT-P-005; IEDT-P-010; IEDT-P-011
Commercial and fandom lexical retention	English nouns and fandom terms index product culture, affective belonging, or lifestyle register	3	20.0	2	2	2	Product, fandom, and identity markers may be over-normalized	"*first luv ku di dunia fangirl*"; "*handcream ... photo set*"; "Penang Food Rocks!"	IEDT-P-002; IEDT-P-003; IEDT-P-006
Orthographic and lexical variant handling	Non-standard spelling, repeated punctuation, or variant form creates ambiguity before translation	2	13.3	2	1	1	Normalization may incorrectly repair intentional or uncertain forms	"*not habits.. tekanan*"; "donut sama doughtnut"	IEDT-P-001; IEDT-P-004
Corrective and rebuttal sequencing	English opening or closing segment shifts stance in an Indonesian argumentative frame	2	13.3	1	2	2	Sequence-sensitive stance may be lost if clauses are processed separately	"No... not like that"; "Well at least some generation can take jokes"	IEDT-P-009; IEDT-P-014
Conversational framing and chat narrative	English discourse marker or chat-related phrasing organizes interactional relation	2	13.3	1	1	2	Interactional stance may be weakened by language-by-language translation	"*Well tergantung*"; "Its kinda weird while you chat"	IEDT-P-013; IEDT-P-015

The dominant pattern is that English insertions often perform the most fragile semantic work. In "*enak not bad lah tapi haram*," the English phrase "not bad" cannot be interpreted apart from Indonesian affective and moral framing. In "server not found," English is not simply a foreign-language phrase but a technical diagnostic expression that should be preserved as a system message. Similarly, "*first luv ku di dunia fangirl*" combines English fandom vocabulary with Indonesian possessive morphology, making full translation less reliable than selective retention. The table shows that affective evaluation carries the highest translation ambiguity and pragmatic loss risk, both appearing in three records. Objectively, this indicates that bilingual meaning is most vulnerable when sentiment, stance, and local discourse particles are distributed across languages.

Meaning preservation in code-switched NLP depends on whether models can distinguish translation from functional retention. A monolingual pipeline may "correct" spelling, translate English phrases, or normalize hybrid expressions, but such repair can remove the very signal that makes the utterance meaningful. This finding resonates with studies arguing that multilingual models are not automatically competent code-switchers and that output alignment remains difficult in code-switched interaction [29], [34], [37], [35]. For this study, normalization is therefore treated as an interpretive risk: computational systems must decide when to standardize, when to translate, and when to preserve mixed-language form.

4.3. Computational risk under monolingual processing

Indonesian–English digital records show that computational risk emerges most sharply when mixed-language utterances are forced into single-language categories. The most frequent risk domain is forced single-language labeling, appearing in five records or 33.3% of this coded corpus. This risk is not simply technical misclassification; it affects how switch-points are recognized, whether English segments are retained, and whether Indonesian matrix frames remain analytically visible. Other risk domains are distributed across sentiment, technical/platform meaning, register, and discourse sequencing. This distribution is consistent with recent code-switched NLP scholarship, which argues that multilingual models may process multiple languages but still fail to represent code-switched interaction as a structurally integrated phenomenon [29], [34], [36].

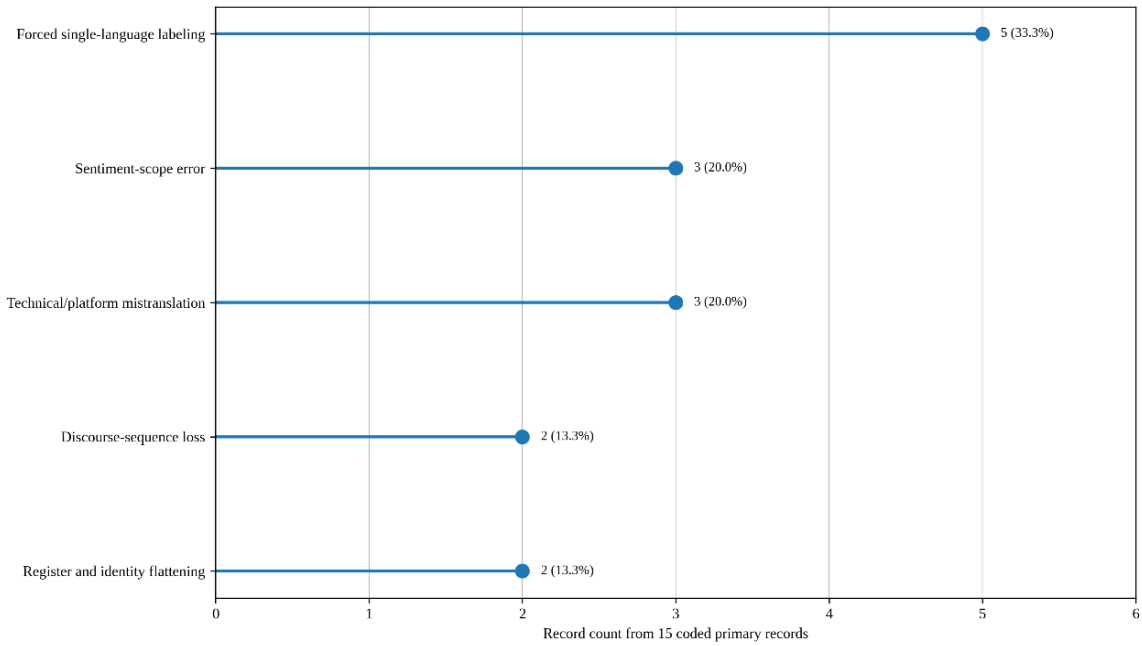

Figure 4. Monolingual processing risk profile in Indonesian–English digital records

Table 4. Monolingual processing risks across Indonesian–English digital records

Risk Domain	Operational Indicator	Frequency	Percentage	Affected Linguistic Zone	Likely Model-Level Failure	Distortion under Monolingual Processing	Representative Excerpt	Data Code
Forced single-language labeling	Mixed utterance is more likely to be assigned to one dominant language despite meaningful alternation	5	33.3	Whole-record language classification; token sequence	Language identification collapse; under-detection of switch-points	English elements become treated as noise, quotation, or removable foreign material	"*first luv ku di dunia fangirl*"; "*respon via DM*"; "*Well tergantung*"	IEDT-P-002; IEDT-P-005; IEDT-P-013
Sentiment-scope error	Evaluative meaning is distributed across Indonesian and English rather than contained in one language segment	3	20.0	Sentiment-bearing phrase; moral or affective clause	Polarity misclassification; incorrect sentiment scope	English polarity is detached from Indonesian affective, moral, or experiential frame	"*enak not bad lah tapi haram*"; "*not so good juga sih*"; "not a good example definitely"	IEDT-P-007; IEDT-P-012; IEDT-P-008
Technical/platform mistranslation	English technical or platform expressions operate as functional labels rather than ordinary lexical items	3	20.0	Platform action; diagnostic phrase; media-sharing template	Overtranslation; domain-term loss; entity-action confusion	Technical or interface meaning is translated literally and loses operational value	"server not found"; "via DM"; "Heres a song for you"	IEDT-P-010; IEDT-P-005; IEDT-P-011
Register and identity flattening	English indexes fandom, lifestyle, commercial, or cosmopolitan identity within Indonesian discourse	2	13.3	Fandom noun; product phrase; lifestyle register	Register loss; weak representation of sociolinguistic stance	Identity-bearing expressions are normalized into generic lexical content	"dunia fangirl"; "*handcream ... photo set*"	IEDT-P-002; IEDT-P-003
Discourse-sequence loss	Stance depends on sequence across clauses, especially when English opens or closes an Indonesian argument	2	13.3	Clause boundary; rebuttal; quoted stance	Clause segmentation error; weakened discourse relation	Correction, irony, or rebuttal becomes fragmented across separate language units	"No... not like that"; "Well at least some generation can take jokes"	IEDT-P-009; IEDT-P-014

The most visible pattern is that monolingual distortion occurs at different layers of language processing. Whole-record labeling is vulnerable in examples such as "*first luv ku di dunia fangirl*," where English lexical items combine with Indonesian possessive morphology and social-domain framing. Sentiment-scope error appears in "*enak not bad lah tapi haram*," where positive English evaluation is constrained by Indonesian moral qualification. Technical mistranslation is visible in "server not found," a phrase whose computational meaning depends on preserving the English diagnostic form. The table also shows that smaller-frequency categories remain analytically important because they involve identity and discourse sequencing. Thus, lower counts do not indicate lower significance; they indicate more specialized but consequential forms of risk.

Monolingual assumptions are inadequate because they treat mixed-language text as a problem of classification before treating it as a problem of meaning. In this corpus, code-switching is not an accidental mixture to be cleaned; it is a mechanism through which stance, technical precision, platform practice, and identity are encoded. This position aligns with Indonesian AI evaluation work that questions translated benchmarks, local-language safety, cultural hallucination, parameter-efficient adaptation, and sociolinguistic bias [9], [12], [13], [14], [11]. For this study, computational adequacy therefore requires preserving multilingual form, identifying switch-points, and interpreting pragmatic function within the same analytic pipeline.

5. DISCUSSION

The first analytical implication is that Indonesian–English code-switching cannot be reduced to a marginal irregularity in otherwise Indonesian digital text. Switch-points function as sites where users compress stance, identity, evaluation, platform practice, and technical reference into short multilingual utterances. This matters because computational systems that treat switching as noise risk removing precisely those elements that make digital talk meaningful. The observed configurations support sociolinguistic accounts that describe code-switching as socially organized practice rather than random alternation [15], [16]. They also extend Indonesian studies that locate switching in education, public communication, social media, and elite discourse [18], [17], [22]. The contribution is therefore methodological as well as conceptual: code-switching should be modeled as communicative structure, not merely as a preprocessing challenge before "real" NLP begins.

The underlying reason for this pattern lies in how Indonesian–English digital talk distributes communicative work across language boundaries. Indonesian frequently provides discourse frame, social deixis, and local pragmatic particles, while English contributes evaluative phrases, technical labels, fandom terms, platform commands, or quotation-like stance. This division does not imply a fixed grammar of switching, but it shows that language alternation is patterned enough to require token-level and sequence-sensitive analysis. Studies on Indonesian–English syntactic configuration and digital code-mixing support this interpretation by showing that switching often follows recognizable structural and functional tendencies [19], [25], [24]. At the same time, computational studies remind us that recognizable patterning does not automatically mean easy modeling [28], [29]. This tension explains why switch-points become both linguistic evidence and computational stress points.

The second implication concerns normalization and translation: these processes are not neutral technical repairs, but interpretive operations that can preserve or erase meaning. In the coded records, fragile meaning often appears where English evaluative phrases, technical expressions, and identity-bearing terms are embedded inside Indonesian discourse frames. This complicates pipelines that treat normalization as spelling correction or translation as semantic equivalence. Research on code-mixed corpora and language identification has already shown that tokenization, normalization, and translation form a connected workflow rather than separate tasks [27], [6]. However, this study pushes that point further by showing that functional retention may be more appropriate than full normalization in some contexts. The implication is practical: computational systems should decide whether to translate, preserve, label, or annotate mixed segments according to discourse function, not only lexical language identity.

This pattern emerges because code-switched meaning is often compositional in form but non-compositional in pragmatic effect. A phrase such as an English evaluation inside an Indonesian moral or affective frame cannot be interpreted by translating each language segment independently. Similarly, technical and platform expressions may be meaningful precisely because they remain in English as interface language, diagnostic label, or circulating digital formula. This helps explain why multilingual NLP work increasingly distinguishes multilingual coverage from code-switching competence [29], [34], [35]. It also clarifies why models trained on standard monolingual or translated benchmarks may underperform when confronted with locally situated multilingual data. Studies on culturally grounded Indonesian LLM

evaluation and multilingual reasoning support this concern [9], [10]. Thus, normalization failure is not only linguistic; it reflects mismatch between benchmark logic and multilingual digital practice. Fawaid et al. [38] likewise compare human intelligence and algorithmic precision in Indonesian translation pedagogy, which underlines that algorithmic handling of Indonesian meaning cannot be taken as self-evidently adequate.

The third implication is that monolingual processing produces representational distortion, not merely classification error. Forced single-language labeling can make English insertions appear peripheral, while Indonesian discourse frames become invisible when English-heavy phrases dominate model attention. Sentiment-scope error, technical mistranslation, register flattening, and discourse-sequence loss are therefore symptoms of a deeper modeling problem: multilingual users are represented through categories that do not match their communicative practice. This finding aligns with recent scholarship on LLM limitations in code-switched understanding, reasoning, and generation [34], [37], [36]. It also resonates with Indonesian AI studies on safety failure, hallucination, and sociolinguistic bias [12], [13], [11]. The larger implication is that model reliability must include representational adequacy, not only aggregate task performance.

The reason monolingual risk persists is structural: many NLP pipelines still begin by assigning language, cleaning input, segmenting text, and optimizing downstream tasks according to assumptions of linguistic separability. Such architecture works best when languages are discrete layers, but Indonesian–English digital talk often uses them as interdependent resources. Multilingual pretraining may reduce vocabulary gaps, yet it does not guarantee sensitivity to switch-point function, local pragmatics, or culturally grounded reference. This explains why parameter-efficient adaptation, local-language evaluation, and bias measurement have become increasingly important in Indonesian computational language research [14], [9], [11]. This study therefore does not claim to prove a universal theory of code-switching; rather, it shows that Indonesian–English digital talk exposes a methodological fault line. Computational adequacy requires models that keep language labels, pragmatic functions, and sociolinguistic representation analytically connected.

6. CONCLUSION

This study concludes that Indonesian–English digital code-switching is not a peripheral irregularity but a structured computational problem that challenges monolingual assumptions in NLP. Its most important lesson is that switch-points carry communicative value: they organize stance, affect, technical meaning, platform practice, and identity within compact digital utterances. By linking token-level language identification, normalization–translation alignment, and pragmatic-computational risk coding, this study renews code-switching research beyond descriptive sociolinguistics and narrow detection tasks. Its contribution lies in reframing mixed-language data as an accountable modeling object, where linguistic structure and computational representation must be examined together.

This study is limited by its text-based corpus design, modest coded record set, and focus on Indonesian–English digital records rather than broader Indonesian multilingual ecologies involving regional languages, speech, images, or multimodal interaction. Its findings should therefore be read as interpretive and diagnostic, not as population-level generalization or causal explanation of user behavior. Future research should expand corpus scale, compare platforms, include regional-language combinations, and test model behavior across tokenization, sentiment classification, translation, and LLM response generation. Further work should also incorporate human annotation reliability and user-centered evaluation to examine whether computational preservation of code-switching improves representational fairness.

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

Devita Septian Dwi Hidayati: 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, anonymized public digital text records.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible digital text records and scholarly materials, 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 digital text records and scholarly materials analyzed in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] DataReportal, *Digital 2025: Indonesia*, 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-indonesia.
[2] J. Santoso, "Indonesian named entity recognition using BiLSTM-CRF with domain-specific word embeddings," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 1-10, 2026.
[3] W. Novitasari, "Sentiment analysis on Indonesian political discourse using transformer-based models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 11-21, 2026.
[4] A. Wahid, "Lexical patterns in Indonesian academic writing: A corpus-based analysis for writing assistance," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 32-42, 2026.
[5] T. Yuliyana, "Automatic readability assessment of Indonesian educational texts using hybrid NLP approaches," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 43-52, 2026.
[6] A. Nugroho, "Code-mixing detection in Indonesian-English tweets using machine learning and linguistic features," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 22-31, 2026.
[7] A. M. Barik, M. Adriani, and R. Mahendra, "Normalization of Indonesian-English code-mixed Twitter data," in *Proc. 5th Workshop on Noisy User-generated Text*, 2019, pp. 417-424, doi: 10.18653/v1/D19-5554.
[8] M. I. Ibrahim, "Morphological segmentation of low-resource Indonesian dialects using unsupervised neural models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 1, pp. 53-62, 2026.
[9] A. E. Bouazzaoui, "Beyond translated benchmarks: A culturally grounded evaluation of large language models for Indonesian language understanding," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[10] F. D. Wulandari, "Do large language models reason consistently across Indonesian languages? A multilingual evaluation of Indonesian, Javanese, Sundanese, and Buginese," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[11] M. A. Sumadi, "Who gets represented by Indonesian AI? Measuring regional, gender, and sociolinguistic bias in large language models," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[12] Nadiyah, "When safety fails in local languages: Evaluating culturally sensitive responses of large language models in multilingual Indonesia," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[13] R. T. Wibisono, "Hallucinating the archipelago: Detecting factual and cultural errors in retrieval-augmented generation for Indonesian knowledge," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[14] A. Hadziqi, "Small models, many languages: Parameter-efficient adaptation of multilingual language models for low-resource Indonesian languages," *Indonesian Journal of Computational Language Studies*, vol. 1, no. 2, 2026.
[15] S. A. Doğruöz, S. Sitaram, B. E. Bullock, and A. J. Toribio, "A survey of code-switching: Linguistic and social perspectives for language technologies," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 1654-1666, doi: 10.18653/v1/2021.acl-long.131.
[16] W. Rayo, A. M. Barrita, L. C. Martinez, and I. Carbajal, "The complexity in bilingual code-switching research: A systematic review," *Languages*, 2024, doi: 10.3390/languages9060217.
[17] A. Azzahra, "Code switching and code mixing in modern public communication: Indonesian sociolinguistic analysis," *International Journal of Current Science Research and Review*, 2024, doi: 10.47191/ijcsrr/v7-i6-58.
[18] W. O. Hanafiah, "Forms and functions of Indonesian-English code switching: A case study of elites in Makassar City," *ELS Journal on Interdisciplinary Studies in Humanities*, 2022, doi: 10.34050/elsjish.v5i3.23155.
[19] H. Sahib, W. Hanafiah, M. Aswad, A. Yassi, and F. Mashhadi, "Syntactic configuration of code-switching between Indonesian and English: Another perspective on code-switching phenomena," *Education Research International*, 2021, doi: 10.1155/2021/3402485.
[20] H. Hafid and M. Margana, "Code-switching practices in multilingual classrooms: Exploring pedagogical functions," *AL-ISHLAH: Jurnal Pendidikan*, 2022, doi: 10.35445/alishlah.v14i2.1326.
[21] S. D. Mulyani, S. Delfi, and J. Jismulatif, "Beyond English and Bahasa: Exploring the pedagogical functions of code-switching in multilingual EFL contexts," *REiLA: Journal of Research and Innovation in Language*, 2024, doi: 10.31849/reila.v6i3.23316.
[22] D. Nuraini and F. Fitriana, "Code-switching as a communicative strategy among Indonesian university students on social media," *Jurnal Tahuri*, 2024, doi: 10.30598/tahurivol21issue2page155-172.
[23] M. J. Yusuf, I. Simpuruh, and S. Syamsinar, "An analysis of code-switching by college students in Makassar," *ELS Journal on Interdisciplinary Studies in Humanities*, 2024, doi: 10.34050/els-jish.v7i4.42024.
[24] M. A. Manuhutu, A. I. Rahmadani, and P. E. I. Suardana, "Language contact phenomena: A case study of Indonesian-English code-switching in social media communication," *Focus Journal Language Review*, 2024, doi: 10.62795/fjlg.v2i2.291.

[25] H. Wibowo and N. Hamidah, "Linguistic interplay on social media: Unraveling Indonesian-English code mixing on Twitter," *Tamaddun*, 2023, doi: 10.33096/tamaddun.v22i2.547.
[26] U. A. S. Tabi'in, "Code-switching in digital communication: A study of Indonesian EFL students on social media platforms," *INOMATEC: Jurnal Inovasi dan Kajian Multidisipliner Kontemporer*, 2026, doi: 10.70294/ino1234.
[27] A. F. Hidayatullah, R. Apong, D. Lai, and A. Qazi, "Corpus creation and language identification for code-mixed Indonesian-Javanese-English tweets," *PeerJ Computer Science*, vol. 9, 2023, doi: 10.7717/peerj-cs.1312.
[28] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, and P. Fung, "Are multilingual models effective in code-switching?," in *Proc. 5th Workshop on Computational Approaches to Linguistic Code-Switching*, 2021, pp. 142-153, doi: 10.18653/v1/2021.calcs-1.20.
[29] R. Zhang, S. Cahyawijaya, J. C. B. Cruz, and A. F. Aji, "Multilingual large language models are not (yet) code-switchers," in *Proc. Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 12567-12582, doi: 10.48550/arxiv.2305.14235.
[30] A. F. Hidayatullah, R. Apong, D. T. C. Lai, and A. Qazi, "Pre-trained language model for code-mixed text in Indonesian, Javanese, and English using transformer," *Social Network Analysis and Mining*, vol. 15, 2025, doi: 10.1007/s13278-025-01444-9.
[31] H. Rathnayake, J. Sumanapala, R. Rukshani, and S. Ranathunga, "Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification," *Knowledge and Information Systems*, vol. 64, pp. 1937-1966, 2022, doi: 10.1007/s10115-022-01698-1.
[32] S. Mondal, R., S. Pathak, P. Jyothi, and A. Raghuveer, "CoCoa: An encoder-decoder model for controllable code-switched generation," in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing*, 2022, pp. 2466-2479, doi: 10.18653/v1/2022.emnlp-main.158.
[33] P. Xie, X. Liu, T. W. Chan, Y. Bie, Y. Song, Y. Wang, H. Chen, and K. Chen, "SwitchLingua: The first large-scale multilingual and multi-ethnic code-switching dataset," arXiv, 2025, doi: 10.48550/arxiv.2506.00087.
[34] A. Mohamed, Y. Zhang, M. Vazirgiannis, and G. Shang, "Lost in the mix: Evaluating LLM understanding of code-switched text," arXiv, 2025, doi: 10.48550/arxiv.2506.14012.
[35] R. Sheth, S. R. Sinha, M. Patil, H. Beniwal, and M. Singh, "Beyond monolingual assumptions: A survey of code-switched NLP in the era of large language models," arXiv, 2025, doi: 10.48550/arxiv.2510.07037.
[36] G. Winata et al., "Can large language models understand, reason about, and generate code-switched text?," arXiv, 2026, doi: 10.48550/arxiv.2601.07153.
[37] J. Oh, H. Yoo, F. G. Haznitrama, and A. Oh, "OLA: Output language alignment in code-switched LLM interactions," arXiv, 2026, doi: 10.48550/arxiv.2601.03589.
[38] 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



Devita Septian Dwi Hidayati is affiliated with Universitas Pembangunan Nasional "Veteran" Jawa Timur, Indonesia, where her work engages computational sociolinguistics, corpus-based discourse analysis, and the study of Indonesian–English code-switching in digital communication. Her current research examines how mixed-language digital talk challenges monolingual assumptions in language identification, normalization, and multilingual NLP representation. She can be contacted at email: 24046010029@student.upnjatim.ac.id