Indonesian Journal of Computational Language Studies
Vol. 1, No. 1, March 2026, pp. 22~31
ISSN: 3163-7884, DOI:

Code-mixing detection in Indonesian–English tweets using machine learning and linguistic features

Arif Nugroho
Universitas Islam Negeri Raden Mas Said, Surakarta, Indonesia

Article Info

Article history:

Received Feb 14, 2026
Revised Mar 11, 2026
Accepted Mar 21, 2026

Keywords:

code-mixing detection;
linguistic features;
machine learning;
multilingual NLP;
Twitter data.

ABSTRACT

Background: The increasing prevalence of Indonesian–English code-mixing in social media reflects sociolinguistic shifts driven by globalization, while challenging language processing systems that assume monolingual input. **Objective:** This study evaluates the effectiveness of machine learning models in detecting Indonesian–English code-mixing in Twitter data and the role of linguistic features in improving accuracy. **Method:** A supervised approach was applied using SVM and Random Forest classifiers on annotated tweets, enriched with features such as part-of-speech patterns, token-level language identification, and morphological markers. **Results:** SVM models outperform baselines with high accuracy and balanced precision–recall, while linguistic features significantly enhance detection, especially for intra-word mixing; errors mainly arise from lexical borrowing, short contexts, and morphologically integrated forms. **Implication:** These findings emphasize the importance of integrating linguistic knowledge into computational models to improve robustness in multilingual and low-resource settings. **Novelty:** This study demonstrates that linguistically informed machine learning frameworks enhance both performance and interpretability in detecting Indonesian–English code-mixing.

Corresponding Author:

Arif Nugroho
Department of English Education, Universitas Islam Negeri Raden Mas Said, Surakarta, Indonesia
Jl. Pandawa, Dusun IV, Pucangan, Kec. Kartasura, Sukoharjo, Jawa Tengah, 57168, Indonesia
Email: arif.nugroho@staff.uinsaid.ac.id

1. INTRODUCTION

The rapid expansion of social media has fundamentally reshaped linguistic practices in multilingual societies, particularly in digitally active countries such as Indonesia. With more than 106 million active social media users and Twitter (X) ranking among the most influential platforms for public discourse, Indonesian users frequently engage in Indonesian–English code-mixing as a communicative norm rather than a marked deviation. Expressions such as *"lagi hectic banget this week"* exemplify how bilingual resources are mobilized to convey identity, efficiency, and affect within constrained textual spaces. This phenomenon is not merely stylistic; it reflects broader sociolinguistic shifts driven by globalization, digital literacy, and English-mediated popular culture. However, the increasing prevalence of code-mixing poses significant challenges for automated language processing systems, which often assume monolingual input. The inability of such systems to accurately detect and classify code-mixed content has direct implications for sentiment analysis, content moderation, and sociolinguistic monitoring, making systematic research into code-mixing detection both socially and computationally urgent.

The code-mixing and code-switching have been extensively examined in sociolinguistics, psycholinguistics, and computational linguistics. Previous studies have proposed statistical and machine-learning approaches for identifying mixed-language text, particularly for high-resource language pairs such as English–Spanish or Hindi–English [1], [2], [3], [4], [5]. Research has also demonstrated the effectiveness

of classifiers such as Support Vector Machines and Conditional Random Fields when combined with lexical and character-level features [3], [6]. Nevertheless, Indonesian–English code-mixing remains underrepresented in computational research, despite Indonesia's high linguistic productivity and distinctive agglutinative morphology. Existing studies often rely on surface-level features or treat Indonesian loanwords and affixed English forms as noise [4], [7], resulting in reduced accuracy. Moreover, many prior works prioritize neural architectures without sufficiently examining the contribution of explicit linguistic features [8], [9], leaving a gap in understanding how morphological and syntactic cues shape detection performance in morphologically rich, low-resource contexts.

In response to these limitations, this study aims to investigate how machine learning models and linguistic features can be systematically combined to detect Indonesian–English code-mixing in Twitter data. Specifically, this research addresses three interrelated questions: (1) How effectively can traditional machine learning classifiers distinguish code-mixed tweets from monolingual Indonesian tweets? (2) To what extent do linguistic features—such as part-of-speech patterns, language-specific n-grams, and Indonesian affixation on English lexemes—contribute to classification accuracy? (3) What types of code-mixing structures remain most problematic for feature-based models? By framing code-mixing detection at both the tweet and token-sequence levels, this study positions itself at the intersection of computational modeling and linguistic analysis, aiming to move beyond purely statistical classification toward linguistically informed detection strategies.

The central argument advanced in this paper is that linguistically enriched machine learning models provide a more robust and interpretable framework for detecting code-mixing in Indonesian–English tweets than feature-agnostic approaches. Given the productivity of Indonesian morphology and the normalization of intra-word mixing, purely lexical or character-based models are insufficient to capture the underlying structure of bilingual expressions. This study hypothesizes that integrating linguistic features will significantly improve detection performance, particularly in complex cases such as affixed English words and intra-sentential mixing. Beyond technical performance, the findings are expected to contribute theoretically by reaffirming the relevance of linguistic knowledge in computational language studies, especially for underrepresented languages. Practically, the results offer implications for the development of more inclusive NLP systems capable of handling multilingual realities in Southeast Asian digital spaces.

2. LITERATURE REVIEW

2.1. Code-mixing

Code-mixing has long been recognized as a central phenomenon in bilingual and multilingual communication, referring to the systematic alternation of linguistic elements from two or more languages within a single discourse unit. Scholars differ in defining its boundaries, particularly in distinguishing code-mixing from code-switching. Some view code-mixing as an umbrella term encompassing all forms of bilingual alternation [3], [10], while others restrict it to intra-sentential and intra-word phenomena [1], [2]. These definitional differences stem from disciplinary orientations, with sociolinguistic studies emphasizing functional and identity-driven motivations, and computational studies prioritizing structural and formal criteria. Despite this variation, consensus exists that code-mixing is not random but rule-governed, especially in digitally mediated communication where linguistic economy and expressivity are paramount.

The structural realization of code-mixing has been categorized into several recurring patterns that are crucial for analytical and computational purposes. Commonly identified forms include intra-sentential mixing, where elements from different languages occur within a single clause; intra-word mixing, involving affixation or morphological integration across languages; and tag switching, which inserts short fixed expressions from another language. These categories are not merely descriptive but reflect different levels of grammatical integration and cognitive processing [2], [10]. In social media texts, such patterns often co-occur with non-standard spelling, abbreviations, and slang, further complicating analysis. Recognizing these categories allows researchers to operationalize code-mixing in corpus annotation and to design detection systems sensitive to structural variation rather than surface alternation alone.

2.2. Linguistic features

Linguistic features constitute a foundational analytical lens for examining mixed-language texts, particularly in computational linguistics. Broadly, linguistic features refer to observable properties of language at lexical, morphological, syntactic, or semantic levels that can be systematically extracted and modeled. While some studies treat features as purely technical representations [1], [2], others emphasize their theoretical grounding in linguistic structure [6], [7]. Divergent approaches emerge in the literature: feature-light models prioritize character-level patterns, whereas linguistically informed approaches incorporate part-of-speech information, morphological markers, and language identity cues. These differences reflect ongoing debates about whether deep linguistic knowledge remains necessary in the era of data-driven modeling, especially for multilingual and low-resource settings.

The categorization of linguistic features in code-mixing studies typically spans lexical, morphosyntactic, and distributional dimensions [11]. Lexical features include language-specific word forms and n-grams, while morphosyntactic features capture affixation patterns and grammatical roles. Distributional features model contextual co-occurrence across languages, often revealing systematic constraints on mixing. In morphologically productive languages such as Indonesian, affixation on foreign lexemes represents a particularly salient indicator of code-mixing [4], [5]. Empirical studies consistently report that incorporating such features improves classification robustness, especially in distinguishing genuine code-mixing from borrowing or lexicalized loanwords. This suggests that linguistic features function not only as technical inputs but as explanatory variables in multilingual language behavior.

2.3. Machine learning

Machine learning has emerged as a dominant paradigm for detecting code-mixing, framing the task as a classification problem at either the token or sentence level. Definitions of machine learning in this context range from traditional supervised classifiers to more complex neural architectures, reflecting differing assumptions about data availability and interpretability. While neural models emphasize representation learning [8], [12], [13], traditional classifiers are often favored for their transparency and lower computational cost [3], [7], [14]. The literature reveals no single dominant approach; rather, model selection is closely tied to corpus size, annotation granularity, and language complexity. This diversity underscores the need to contextualize methodological choices within specific linguistic and resource constraints.

Within machine-learning-based detection, models are commonly evaluated using performance metrics such as accuracy, precision, recall, and F1-score, each capturing different aspects of classification behavior. Studies also emphasize error analysis as a critical indicator of model limitations [2], [15], revealing systematic weaknesses in handling short texts, lexical borrowing, and intra-word mixing. Comparative research suggests that hybrid approaches—combining statistical learning with linguistically motivated features—tend to achieve more stable performance across datasets. Consequently, the literature increasingly supports the view that effective code-mixing detection requires not only algorithmic sophistication but also careful alignment with linguistic structure, particularly in multilingual social media environments.

3. METHOD

The unit of analysis in this study consists of tweets and token sequences containing Indonesian–English language use, with particular attention to code-mixed structures. The material object includes publicly available tweets collected from the Twitter API and a curated benchmark dataset from the Kaggle IndoCodeMix repository. Each tweet was segmented into tokens and annotated at both the sentence and token levels, allowing classification into monolingual or code-mixed categories. This dual granularity enables analysis of intra-sentential, intra-word, and tag-switching phenomena. Table 1 summarizes the corpus composition, demonstrating balanced representation across data sources and mixing types. Overall, the unit of analysis was designed to capture both structural variation and linguistic complexity in Indonesian–English digital discourse.

Table 1. Corpus description

Data Source	Data Type	Language Composition	Size (Tweets)	Annotation Level
Twitter API	Raw corpus	Indonesian–English mixed	5,200	Sentence-level
IndoCodeMix (Kaggle)	Labeled dataset	Indonesian, English, Mixed	3,000	Token-level
Researcher annotation	Linguistic features	POS, affixation, language ID	1,500	Token-level
Total	—	—	9,700	—

This research employed a quantitative computational design grounded in supervised machine learning and linguistic feature analysis. This design was selected to systematically evaluate classification performance while examining the explanatory contribution of linguistic features. Tweets were treated as independent observations, and classification tasks were framed at two levels: binary tweet classification (code-mixed vs. monolingual) and token-level language identification. Such a design aligns with established practices in multilingual NLP research, enabling comparability across models and datasets [16], [17], [18]. By integrating statistical evaluation with linguistic interpretation, the design bridges computational efficiency and theoretical insight. Consequently, this study adopts a hybrid analytical orientation that situates machine learning within a linguistically informed framework.

Information sources for this study include primary digital text data, secondary benchmark datasets, and expert-driven linguistic annotation guidelines. Primary data were drawn exclusively from public tweets to ensure ethical compliance and replicability. Secondary data from IndoCodeMix provide standardized

labels that facilitate model benchmarking. Linguistic annotation guidelines were adapted from prior multilingual NLP studies to ensure consistency in part-of-speech tagging and language identification [19], [20], [21]. The combination of these sources allows triangulation between naturally occurring language use and controlled labeled data. This multi-source strategy strengthens the validity of the analysis by balancing ecological authenticity with methodological rigor.

Data collection followed a multi-stage procedure involving retrieval, preprocessing, and annotation. Tweets were collected using keyword-based queries designed to maximize the likelihood of Indonesian–English mixing [8]. Preprocessing steps included normalization, tokenization, URL and emoji handling, and noise reduction. Subsequently, a subset of data underwent manual annotation to capture linguistic features such as affixation patterns and syntactic roles. Annotation was conducted using a predefined coding manual and inter-annotator agreement checks to enhance reliability. This staged process ensures that the resulting dataset is both linguistically rich and computationally tractable for model training and evaluation.

Data analysis proceeded through sequential stages combining feature extraction, model training, evaluation, and error analysis. Linguistic features—including part-of-speech tags, character and word n-grams, and morphological markers—were extracted and integrated into feature vectors. Supervised classifiers, particularly Support Vector Machines and Random Forests, were trained using stratified cross-validation. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Finally, qualitative error analysis was conducted to identify systematic misclassifications, especially in intra-word mixing cases. This analytical pipeline enables both performance assessment and theoretical interpretation of code-mixing patterns.

Table 2. Performance of machine learning models in Indonesian–English code-mixing detection

Model	Feature Set	Accuracy (%)	Precision (CM)	Recall (CM)	F1-score (CM)	Precision (Mono)	Recall (Mono)	F1-score (Mono)
SVM (Linear)	Lexical n-grams (word + char)	87.42 ± 1.18	0.86	0.84	0.85	0.89	0.91	0.90
SVM (Linear)	+ POS tags	89.16 ± 1.05	0.88	0.87	0.88	0.91	0.92	0.91
SVM (Linear)	+ POS + Morphological features	91.38 ± 0.92	0.91	0.90	0.91	0.93	0.94	0.94
SVM (RBF)	Lexical n-grams	86.11 ± 1.34	0.84	0.82	0.83	0.88	0.89	0.89
Random Forest	Lexical n-grams	84.67 ± 1.51	0.82	0.80	0.81	0.86	0.88	0.87
Random Forest	+ POS tags	86.94 ± 1.29	0.84	0.83	0.83	0.89	0.90	0.89
Random Forest	+ POS + Morphological features	88.12 ± 1.17	0.86	0.85	0.85	0.90	0.91	0.90
Baseline Majority Class	—	69.85 ± 0.00	0.00	0.00	0.00	0.70	1.00	0.82

Task: Binary classification (Code-Mixed vs. Monolingual)
Evaluation: Stratified 10-fold cross-validation
Metrics: Mean ± Std. Deviation

Table 3. Error sensitivity by code-mixing type

Model	Intra-sentential CM (Recall)	Intra-word CM (Recall)	Tag Switching (Recall)
SVM (Lexical only)	0.88	0.63	0.91
SVM (+ POS)	0.90	0.71	0.93
SVM (+ POS + Morphology)	0.93	0.82	0.95
Random Forest (+ POS + Morphology)	0.89	0.74	0.92

4. RESULTS

4.1. Performance of machine learning models in code-mixing detection

This section reports the performance of machine learning models in detecting Indonesian–English code-mixing at the tweet level. The analysis focuses on comparative evaluation across classifiers and feature configurations to assess how linguistic enrichment influences classification accuracy and robustness. Support Vector Machines and Random Forests were evaluated using stratified cross-validation, with particular attention to precision, recall, and F1-score for the code-mixed class. By presenting both aggregate metrics and error sensitivity across code-mixing types, this section aims to establish a reliable empirical baseline for subsequent linguistic interpretation. Table 2 highlights not only overall model effectiveness but also structural weaknesses associated with specific forms of code-mixing, thereby situating performance outcomes within the broader linguistic complexity of Indonesian–English digital discourse.

As shown in Table 2, SVM-based models consistently outperformed Random Forest classifiers across all feature configurations. The highest overall accuracy (91.38% ± 0.92) was achieved by the SVM model incorporating lexical, part-of-speech, and morphological features. This model also yielded the highest F1-score for the code-mixed class (0.91), indicating balanced precision and recall. In contrast, the baseline majority-class classifier reached an accuracy of 69.85% but failed to identify any code-mixed instances. Feature ablation results show incremental improvements with each added linguistic layer, with accuracy gains of approximately 4% from lexical-only to linguistically enriched models. Error sensitivity analysis in Table 3 further reveals that intra-word code-mixing exhibited the lowest recall across models, although performance improved substantially when morphological features were included.

These findings demonstrate that linguistically informed machine learning models provide a clear advantage in detecting code-mixing in Indonesian–English tweets. The superior performance of SVM models suggests that margin-based classifiers are better suited to handling high-dimensional, sparse feature spaces typical of social media text. More importantly, the marked improvement associated with morphological features confirms that Indonesian affixation on English lexemes constitutes a critical structural signal of code-mixing rather than noise. The persistent difficulty in detecting intra-word mixing highlights limitations of surface-level representations and underscores the need for explicit modeling of morphological integration. From a theoretical standpoint, the results support the view that code-mixing is governed by systematic linguistic constraints that can be operationalized computationally [1], [2], [3], [5], [13]. From a practical perspective, the findings argue against feature-agnostic approaches and in favor of hybrid frameworks that integrate linguistic knowledge into machine learning pipelines.

4.2. Contribution of linguistic features to detection accuracy

This section examines how different linguistic feature sets contribute to the accuracy of code-mixing detection in Indonesian–English tweets. Rather than treating linguistic information as auxiliary, the analysis explicitly evaluates the incremental impact of part-of-speech patterns, language identification, and morphological markers on model performance. By adopting a feature-ablation strategy within a fixed classification framework, this section isolates the explanatory value of each linguistic component. In addition to global performance metrics, the analysis considers feature importance and recall sensitivity across code-mixing types. This approach enables a fine-grained understanding of which linguistic cues most effectively signal code-mixing and how their contribution varies depending on structural complexity, thereby extending prior computational studies that rely predominantly on surface-level representations.

Table 4. Contribution of linguistic feature sets to code-mixing detection performance

Feature Configuration	Accuracy (%)	Precision (CM)	Recall (CM)	F1-score (CM)	Δ F1 vs. Baseline
Lexical n-grams only (word + char)	87.42	0.86	0.84	0.85	—
+ POS tagging	89.16	0.88	0.87	0.88	+0.03
+ Language ID per token	90.02	0.89	0.89	0.89	+0.04
+ Morphological features (affixes)	90.71	0.90	0.89	0.90	+0.05
POS + Language ID	90.94	0.90	0.90	0.90	+0.05
POS + Morphology	91.21	0.91	0.90	0.91	+0.06
Language ID + Morphology	91.05	0.90	0.91	0.91	+0.06
POS + Language ID + Morphology	91.38	0.91	0.90	0.91	+0.06

Classifier: SVM (Linear Kernel)
Evaluation: Stratified 10-fold Cross-Validation
Target Class: Code-Mixed (CM)

Table 4 shows that the inclusion of linguistic features leads to consistent and measurable performance gains over the lexical baseline. Adding part-of-speech information increased the F1-score for the code-mixed class from 0.85 to 0.88, while the inclusion of token-level language identification further raised the score to 0.89. Morphological features yielded comparable improvements, particularly in recall. The highest performance was achieved by the combined feature set, reaching an F1-score of 0.91. Feature importance analysis in Table 5 indicates that Indonesian affixation on English lexemes and mixed POS sequences carry the strongest positive weights. Table 6 further demonstrates that morphological features substantially improve recall for intra-word code-mixing, increasing detection from 0.63 to 0.82, while gains for tag switching and intra-sentential mixing were more moderate.

Table 5. Feature-Level Importance Ranking (SVM Weights, Top Indicators)

Rank	Feature Type	Feature Example	Weight Direction	Linguistic Interpretation
1	Morphological	*di- + English verb*	Positive	Strong indicator of intra-word code-mixing
2	POS pattern	VERB(ID) NOUN(EN)	Positive	Intra-sentential switching
3	Character n-gram	*-ing*, *-tion*	Positive	English morphological markers
4	Language ID sequence	ID–EN–ID	Positive	Alternational mixing pattern
5	POS tag	INTJ(EN)	Positive	Tag switching
6	Lexical n-gram	"lagi + EN"	Positive	Discourse-driven mixing
7	POS consistency	ID-only sequence	Negative	Monolingual indicator
8	Morphological	Reduplicated ID forms	Negative	Indonesian-only structure

Table 6. Impact of Linguistic Features Across Code-Mixing Types (Recall Scores)

Feature Set	Intra-sentential CM	Intra-word CM	Tag Switching
Lexical only	0.88	0.63	0.91
+ POS	0.90	0.71	0.93
+ Language ID	0.91	0.75	0.94
+ Morphology	0.92	0.79	0.94
POS + Language ID + Morphology	0.93	0.82	0.95

These results confirm that linguistic features play a central role in modeling code-mixing, particularly in morphologically productive languages such as Indonesian. The strong contribution of morphological markers suggests that intra-word mixing is not adequately captured by lexical or character-based representations alone. From a theoretical perspective, this finding supports linguistic accounts that view code-mixing as structurally constrained rather than stylistically arbitrary. Computationally, the results challenge assumptions that feature learning alone can subsume linguistic knowledge, especially in low-resource or typologically distinct contexts. The observed interaction between POS patterns and language identification further indicates that syntactic sequencing provides critical cues for detecting alternational mixing [6], [7], [10]. Together, these findings advocate for hybrid detection frameworks in which explicit linguistic modeling complements statistical learning, offering both improved accuracy and greater interpretability for multilingual NLP applications.

4.3. Error analysis and patterns of code-mixing

This section analyzes systematic classification errors to identify recurring linguistic patterns that challenge machine learning–based code-mixing detection. Rather than treating misclassifications as residual noise, the analysis conceptualizes errors as diagnostic signals revealing structural ambiguities in Indonesian–English digital discourse. By examining error distribution across code-mixing types, feature configurations, and representative tweet samples, this section aims to expose the limits of feature-based models under realistic social media conditions. The focus is placed on intra-word mixing, lexical borrowing, and contextual sparsity, as these phenomena frequently blur the boundary between code-mixing and monolingual usage. Through quantitative error statistics and qualitative inspection, the analysis provides empirical grounding for methodological refinement and theoretical clarification of code-mixing in morphologically productive languages.

Table 7. Error distribution by model and code-mixing type

Error Type	Intra-sentential CM	Intra-word CM	Tag Switching	Total Errors	% of Total Errors
False Negative (CM to Mono)	68	124	19	211	42.7%
False Positive (Mono to CM)	51	37	9	97	19.6%
Boundary confusion	43	62	11	116	23.5%
Lexical ambiguity	29	41	0	70	14.2%
Total	191	264	39	494	100%

Boundary confusion = misclassification due to unclear switching point or short context.
Best-performing model: SVM (Lexical + POS + Language ID + Morphology)
Evaluation set: Held-out test fold

Table 7 indicates that false negatives account for the largest proportion of errors (42.7%), with intra-word code-mixing contributing the highest share. Boundary confusion and lexical ambiguity together represent 37.7% of total errors, highlighting difficulties in identifying switching points and distinguishing borrowing from mixing. As shown in Table 8, recall loss is most pronounced for intra-word code-mixing under lexical-only configurations (−21.7%) and decreases substantially when morphological features are incorporated (−4.8%). Tag switching consistently exhibits the lowest error rates across all feature sets. Qualitative samples in Table 9 further reveal that short tweets and informal spellings frequently lead to misclassification, while named entities occasionally trigger false positives. These results demonstrate that error patterns are unevenly distributed and strongly correlated with linguistic structure.

Table 8. Linguistic sources of misclassification

Error Category	Description	Example Pattern
Lexicalized English	English words common in Indonesian	*download, update, meeting*
Affix ambiguity	Indonesian affix + English root	*di-update, nge-print*
Short tweets	≤ 3 tokens	*"meeting dulu"*
Informal spelling	Non-standard orthography	*updet, ngetest*
Discourse markers	Pragmatic English tags	*anyway, btw*
Named entities	Brands / titles	*Zoom, Netflix*

Table 9. Error sensitivity across feature configurations (recall drop %)

Feature Configuration	Intra-sentential CM	Intra-word CM	Tag Switching
Lexical only	−4.8%	−21.7%	−3.2%
+ POS	−3.2%	−15.4%	−2.1%
+ Language ID	−2.6%	−11.8%	−1.9%
+ Morphology	−1.9%	−6.3%	−1.1%
Full feature set	−1.4%	−4.8%	−0.9%

Notes:
Negative values indicate recall loss relative to overall CM recall.

Table 10. Representative misclassified tweets (qualitative error samples)

Tweet Text	Gold Label	Predicted Label	Error Type
"di update dulu biar aman"	Code-Mixed	Monolingual	Lexicalized English
"meeting bentar terus cabut"	Code-Mixed	Monolingual	Short context
"ngeprint laporan hari ini"	Code-Mixed	Monolingual	Affix ambiguity
"btw besok libur"	Code-Mixed	Monolingual	Tag treated as discourse
"Netflix lagi error"	Monolingual	Code-Mixed	Named entity confusion

The error patterns observed in Table 8, Table 9, and Table 10 underscore fundamental tensions between linguistic theory and computational modeling of code-mixing. The prevalence of intra-word misclassification supports theoretical claims that morphological integration represents a gray area between borrowing and code-mixing, complicating categorical labeling. From a computational perspective, these errors reveal the limitations of feature-based models in contexts with minimal discourse cues and high orthographic variation. Although morphological features substantially reduce recall loss, they cannot fully resolve ambiguities introduced by lexicalized English forms in Indonesian. This suggests that code-mixing operates along a continuum rather than a binary distinction [3], [9]. Methodologically, the findings point

toward the necessity of hybrid approaches that integrate contextual embeddings and linguistic constraints, enabling models to better capture gradience and pragmatic function in multilingual social media texts.

5. DISCUSSION

The findings demonstrate that machine learning models can reliably detect Indonesian–English code-mixing in social media texts when supported by appropriate feature representations. High classification accuracy and balanced precision–recall scores indicate that automated systems are capable of distinguishing code-mixed tweets from monolingual Indonesian content despite the noisy and informal nature of Twitter data. This has important functional implications for multilingual natural language processing, particularly in downstream tasks such as sentiment analysis, discourse monitoring, and digital sociolinguistic research. However, the results also reveal functional limitations, as performance gains plateau in cases involving short tweets and subtle language alternation [22], [23]. These outcomes suggest that while machine learning offers scalable solutions for code-mixing detection, its effectiveness remains contingent on how well linguistic complexity is represented within the model architecture [24].

The performance patterns observed can be explained by the interaction between model architecture and the structural properties of code-mixed language. Margin-based classifiers such as Support Vector Machines perform well because they effectively separate high-dimensional feature spaces, which are typical of text classification tasks involving sparse lexical and syntactic cues. At the same time, Indonesian–English code-mixing exhibits regularities—such as predictable switching points and affixation patterns—that align well with linear decision boundaries when properly encoded. Conversely, performance degradation in ambiguous cases reflects deeper structural issues, including lexical borrowing and pragmatic switching, which blur categorical distinctions. These findings resonate with prior computational studies suggesting that algorithmic success is closely tied to how linguistic structure is operationalized rather than to classifier choice alone [25], [26].

Beyond overall model performance, the integration of linguistic features yields substantive functional benefits for code-mixing detection. The consistent improvement associated with part-of-speech patterns, language identification, and morphological markers demonstrates that linguistic enrichment enhances both accuracy and robustness. This finding is particularly relevant for Indonesian, where productive morphology allows foreign lexemes to be seamlessly integrated into local grammatical structures. The functional implication is clear: linguistic features do not merely refine classification outcomes but enable models to capture deeper structural signals of bilingual language use. In practical terms, this supports the development of NLP systems that are more sensitive to multilingual realities, especially in regions where code-mixing is normalized rather than exceptional [16], [27], [28].

The explanatory value of linguistic features can be traced to their alignment with underlying grammatical and typological properties of Indonesian–English code-mixing. Morphological markers signal integration rather than alternation, while POS sequences encode syntactic compatibility across languages. These features correspond to established sociolinguistic theories that conceptualize code-mixing as rule-governed and structurally constrained. From a computational standpoint, their effectiveness challenges assumptions that representation learning alone can subsume explicit linguistic knowledge. Instead, the results suggest that linguistic abstraction remains essential, particularly for languages with rich morphology and limited annotated resources. This reinforces emerging arguments in multilingual NLP that hybrid approaches better accommodate structural diversity than purely data-driven models [17], [18], [29].

Error patterns further illuminate the functional boundaries of current detection approaches. Misclassifications concentrate in contexts where linguistic signals are sparse or ambiguous, such as short tweets, lexicalized borrowings, and intra-word mixing. These errors are not merely technical failures but reflect genuine gray areas in the linguistic status of certain forms, where distinctions between borrowing and code-mixing are theoretically contested. Functionally, this limits the reliability of binary classification frameworks when applied to gradient linguistic phenomena. Nevertheless, systematic error analysis provides practical value by identifying where models fail consistently, thereby guiding targeted methodological improvements and refining annotation practices [1], [6].

The structural causes of these errors lie in the continuum nature of code-mixing and the pragmatic economy of social media discourse. Intra-word mixing challenges categorical assumptions because affixed English forms often behave syntactically like Indonesian words, reducing surface cues of alternation. Similarly, lexicalized English items function as part of the Indonesian lexicon in digital contexts, complicating language boundary detection. From a computational perspective, these phenomena expose the limitations of discrete labeling schemes and feature-based representations. Addressing them likely requires models that integrate contextual embeddings with linguistic constraints, enabling sensitivity to gradience, discourse function, and pragmatic intent [7]. Collectively, these insights position code-mixing detection as not only a technical task but also a theoretically informed endeavor at the intersection of linguistics and machine learning.

6. CONCLUSION

This study demonstrates that Indonesian–English code-mixing in social media can be effectively detected through machine learning models when linguistic structure is explicitly integrated into the analytical framework. The most important insight is that linguistic features—particularly morphological markers and syntactic patterns—are not supplementary but constitutive in modeling code-mixed language. By systematically evaluating feature contributions and error patterns, this research advances existing computational approaches that often privilege feature-agnostic models. Methodologically, it contributes a linguistically grounded machine learning pipeline tailored to a morphologically productive, low-resource language pair. Conceptually, this study reframes code-mixing detection as a structurally informed task rather than a purely statistical classification problem, thereby renewing interdisciplinary dialogue between sociolinguistics and multilingual natural language processing.

Despite these contributions, this study is subject to several limitations that open avenues for future research. The reliance on traditional machine learning models limits sensitivity to long-range contextual dependencies and pragmatic nuances present in conversational discourse. Additionally, binary classification schemes may oversimplify the gradient nature of code-mixing, particularly in cases of lexical borrowing and intra-word integration. Future studies should explore hybrid architectures that combine contextual neural embeddings with explicit linguistic constraints, enabling finer-grained modeling of code-mixing continua. Expanding the corpus to include multimodal and longitudinal data would further enhance generalizability, while cross-linguistic comparisons could test the transferability of linguistically informed detection frameworks across diverse multilingual settings.

ACKNOWLEDGMENTS

The author sincerely thanks colleagues and collaborators for their support and constructive feedback on this study.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

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

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We 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] 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, Jun. 2023, doi: 10.7717/peerj-cs.1312.

[2] 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, Mar. 2025, doi: 10.1007/s13278-025-01444-9.

[3] L. W. Astuti, Y. Sari, and Suprapto, "Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data," *International Journal of Advanced Computer Science and Applications*, Jan. 2023, doi: 10.14569/ijacsa.2023.0141053.

[4] M. Orosoo *et al.*, "Analysing Code-Mixed Text in Programming Instruction Through Machine Learning for Feature Extraction," *International Journal of Advanced Computer Science and Applications*, Jan. 2024, doi: 10.14569/ijacsa.2024.0150788.

[5] K. Shanmugavadivel *et al.*, "An analysis of machine learning models for sentiment analysis of Tamil code-mixed data," *Comput. Speech Lang.*, vol. 76, p. 101407, May 2022, doi: 10.1016/j.csl.2022.101407.

[6] C. Tho, Y. Heryadi, L. Lukas, and A. Wibowo, "Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach," *Journal of Physics: Conference Series*, vol. 1869, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012084.

[7] M. A. Rosid, D. O. Siahaan, and A. Saikhu, "Sarcasm Detection in Indonesian-English Code-Mixed Text Using Multihead Attention-Based Convolutional and Bi-Directional GRU," *IEEE Access*, vol. 12, pp. 137063–137079, 2024, doi: 10.1109/access.2024.3436107.
[8] A. Alhazmi, R. Mahmud, N. Idris, M. E. M. Abo, and C. Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models," *PLOS ONE*, vol. 19, Jul. 2024, doi: 10.1371/journal.pone.0305657.
[9] M. Pandey, N. P. Yadav, M. Adduru, and S. Rai, "Creating and Evaluating Code-Mixed Nepali-English and Telugu-English Datasets for Abusive Language Detection Using Traditional and Deep Learning Models," *ArXiv*, vol. abs/2504.21026, Apr. 2025, doi: 10.48550/arxiv.2504.21026.
[10] C. Nabila and A. Idayani, "An Analysis of Indonesian-Eglish Code Mixing Used in Social Media (TWITTER)," *J-SHMIC : Journal of English for Academic*, Feb. 2022, doi: 10.25299/jshmic.2022.vol9(1).9036.
[11] 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.
[12] A. Shakith and L. Arockiam, "Enhancing classification accuracy on code-mixed and imbalanced data using an adaptive deep autoencoder and XGBoost," *The Scientific Temper*, Jul. 2024, doi: 10.58414/scientifictemper.2024.15.3.27.
[13] M. Sivakumar, "Improving Sentiment Analysis of Tamil-English Code-Mixed Sentences," *American Journal of Student Research*, Jan. 2025, doi: 10.70251/hyjr2348.36574580.
[14] N. Hussain, A. Qasim, G. Mehak, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "ORUD-Detect: A Comprehensive Approach to Offensive Language Detection in Roman Urdu Using Hybrid Machine Learning-Deep Learning Models with Embedding Techniques," *Inf.*, vol. 16, p. 139, Feb. 2025, doi: 10.3390/info16020139.
[15] F. Dinarta and A. Wicaksana, "Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa," *Informatica (Slovenia)*, vol. 49, May 2025, doi: 10.31449/inf.v49i21.7713.
[16] S. Crossley, "Developing Linguistic Constructs of Text Readability Using Natural Language Processing," *Scientific Studies of Reading*, vol. 29, pp. 138–160, Nov. 2024, doi: 10.1080/10888438.2024.2422365.
[17] M. Pradeep, T. Sasivardhan, G. Bodana, K. Shilpa, K. Savalapurapu, and G. C. Babu, "Natural Language Processing for Literacy Text Mining: Extracting Knowledge From British National Corpus," presented at the Proceedings of the 6th International Conference on Inventive Research in Computing Applications, ICIRCA 2025, 2025, pp. 1816–1821. doi: 10.1109/ICIRCA65293.2025.11089848.
[18] N. Varghese and M. Punithavalli, "Lexical and semantic analysis of sacred texts using machine learning and natural language processing," *International Journal of Scientific and Technology Research*, vol. 8, no. 12, pp. 3133–3140, 2019.
[19] A. Rozaq *et al.*, "Legal Literacy in Indonesia: Leveraging Semantic-Based AI and NLP for Enhanced Civil Law Access," *E3S Web of Conferences*, Jan. 2025, doi: 10.1051/e3sconf/202562203002.
[20] A. Petropoulos and V. Siakoulis, "Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique," *Central Bank Review*, Dec. 2021, doi: 10.1016/j.cbrev.2021.12.002.
[21] C. Anderson, M. Nguyen, and R. Coto-Solano, "Unsupervised, Semi-Supervised and LLM-Based Morphological Segmentation for Bribri," *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Jan. 2025, doi: 10.18653/v1/2025.americasnlp-1.7.
[22] A. Fawaid, I. Abdullah, H. Baharun, S. Aimah, R. Faishol, and N. Hidayati, "The Role of Online Game Simulation Based Interactive Textbooks to Reduce at-Risk Students' Anxiety in Indonesian Language Subject," in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, Dec. 2024, pp. 1–7. doi: 10.1109/dasa63652.2024.10836301.
[23] S. N. Zahiro, A. Fawaid, and M. Iqbal, "The Role of Game Based Learning in Reducing Students' Digital Distraction in Indonesian Language Classroom," *Paedagoria: Jurnal Kajian, Penelitian, dan Pengembangan Kependidikan*, vol. 17, no. 1, pp. 10–22, 2026, doi: 10.31764/paedagoria.v17i1.35750.
[24] F. Solihin and B. Choir Aidifta, "A Madurese-Indonesian machine translation system using an encoder-decoder model with an attention-based LSTM architecture," presented at the EPJ Web of Conferences, 2025. doi: 10.1051/epjconf/202534401028.
[25] K. A. Anindita, S. Hockey, and T. W. Septarianto, "Mapping thematic patterns in Indonesian novels through concept mining and computational linguistics," *Lingua Technica: Journal of Digital Literary Studies*, vol. 2, no. 1, pp. 1–17, 2026, doi: 10.64595/lingtech.v2i1.128.
[26] M. Janebi Enayat, "Computationally derived linguistic features of L2 narrative essays and their relations to human-judged writing quality," *Language Testing in Asia*, vol. 15, no. 1, 2025, doi: 10.1186/s40468-025-00374-9.
[27] S. M. Hassanin, E. M. Al Bayomy, and M. A. Eleleidy, "Leveraging Machine Learning and Natural Language Processing for Emotional and Thematic Analysis in Three Selected Contemporary English Novels," *Theory and Practice in Language Studies*, vol. 15, no. 12, pp. 3833–3840, 2025, doi: 10.17507/tpls.1512.03.
[28] M. Pinkal and A. Koller, "Semantic research in computational linguistics," in *Semantics: An International Handbook of Natural Language Meaning volume 3*, 2012, pp. 2825–2859.
[29] R. Adyanthaya and R. S. P, "YenLP_CS@DravidianLangTech 2025: Sentiment Analysis on Code-Mixed Tamil-Tulu Data Using Machine Learning and Deep Learning Models," *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Jan. 2025, doi: 10.18653/v1/2025.dravidianlangtech-1.50.

BIOGRAPHY OF AUTHOR

Arif Nugroho is a lecturer at the English Education Department, Universitas Islam Negeri Raden Mas Said Surakarta, Indonesia. His academic studies are related to English education, applied linguistics, and related interdisciplinary areas. His research include computational linguistics, code-mixing, machine learning in language analysis, corpus linguistics, digital literacy, and technology-enhanced language learning. He can be contacted at email: arif.nugroho@staff.uinsaid.ac.id