Attaqwa: Jurnal Ilmu Pendidikan Islam
Vol. 21, Issue (1) 2025, pp. 69-81
p-ISSN 1693-0649 e-ISSN: 2620-3901, DOI: https://doi.org/10.54069/attaqwa.v21i1.978

.

Bridging the Lingustic Gap: Challenges in Building AI Models For Non-Standard Dialects

Rizky Surya Ramadhan [1], Nurul Azizah Ria Kusrini[2] Ardianto[3]

[1] Universitas KH. Abdul Chalim: ramadhanrizkysurya4@gmail.com
[2] Universitas KH. Abdul Chalim: nurulazizah968@gmail.com
[3] Institut Agama Islam Daruttaqwa Gresik: ardianto@insida.ac.id

ARTICLE INFO

Keywords:

Non-standard Indonesian;
Low-resource dialects;
Code-mixing;
Natural Language Processing (NLP);
Language model robustness.

Article history:

Received 2025-06-12
Revised 2025-07-08
Accepted 2025-09-02

ABSTRACT

This study examines the challenges of developing Natural Language Processing (NLP) models for non-standard and low-resource Indonesian dialects, with a focus on code-mixing, slang, and regional variations commonly encountered in digital communication. Using a synthetic dataset (NusaDialect benchmark) for sentiment analysis and Named Entity Recognition (NER), we examined the performance of widely used models, including mBERT, IndoBERT, XLM-RoBERTa, and GPT-4. Quantitative results reveal a significant performance gap when models trained on standard Indonesian are applied to dialectal input, with IndoBERT outperforming mBERT but being surpassed by XLM-RoBERTa. In contrast, GPT-4 demonstrates strong resilience in zero-shot settings. Qualitative error analysis further reveals systematic weaknesses related to out-of-vocabulary slang, code-switching ambiguity, morphological complexity, and pragmatic or culturally embedded expressions. To address these limitations, two mitigation strategies were tested: continued pretraining on social media data and data augmentation with back-translation. Findings indicate that while continued pretraining yields the most significant performance gains, augmentation offers a more balanced trade-off by improving dialectal robustness without degrading performance on formal Indonesian. The study concludes that overcoming these linguistic challenges requires not only technical solutions but also culturally informed approaches. Practical implications extend to AI applications in customer service, social media analysis, and digital governance, where inclusivity and accessibility for diverse language users are essential.

1. INTRODUCTION

The rapid growth of artificial intelligence (AI), particularly Large Language Models (LLMs) such as GPT-4, Claude, and Gemini, has transformed digital communication, translation, and knowledge access. However, this progress is uneven (Kurniawan et al., 2024; Sain et al., 2025). While high-resource languages like English and Mandarin benefit from highly accurate AI tools, speakers of low-resource languages and non-standard dialects face significant barriers (Joshi et al., 2021). This imbalance creates digital exclusion and risks deepening linguistic inequality on a global scale.

Indonesia illustrates this problem sharply. As the fourth most populous nation, it is home to over 700 languages and dialects, with Bahasa Indonesia serving as a second language for the majority of its population (Nurhalisa et al., 2025; Topuha et al., 2025; Wulandari et al., 2025; Zein, 2020). Everyday communication in Indonesia departs significantly from the formal standard: regional slang, colloquialisms, and frequent code-mixing with English and local languages dominate digital platforms. While such practices are culturally natural, they pose challenges for AI models trained primarily on standardized corpora, which often misinterpret or fail to process these forms (Elice et al., 2025; Sodikin, 2024; Syukur et al., 2024).

Previous research highlights why these issues persist. Blodgett et al., (2020) showed how linguistic bias emerges when models privilege standard varieties, while (Arif et al., 2025; Joshi et al., 2021; Khotimah et al., 2024; Reksiana et al., 2024; Rohmiati, 2025) emphasized the scarcity of annotated corpora for low-resource languages. García & Wei, (2014) reframed code-mixing as translanguaging, demonstrating its sophistication as a communicative practice. Empirical projects have begun filling resource gaps—BibleNLP leveraged religious translations (Kohler, 2019), Winata et al., (2019) proposed embeddings for code-mixed language, Nusantara NLP constructed benchmarks for Indonesian languages (Cahyawijaya et al., 2023), and Abdalla et al., (2023) confirmed the persistent underperformance of LLMs on low-resource languages. Yet, most of this work targets either local languages in isolation or formal Bahasa Indonesia, leaving non-standard dialects and informal, code-mixed varieties largely unaddressed.

This study addresses the research gap by conducting a qualitative error analysis of NLP models applied to non-standard Indonesian. Instead of comparing numerical benchmarks, the research aims to classify and interpret the types of errors models make when processing dialectal, colloquial, and code-mixed text. The novelty of this approach lies in its focus on *why* models fail rather than *how much* they fail. The objective of this study is to explore and interpret the types of errors that NLP models produce when processing non-standard Indonesian dialects. By conducting a qualitative error analysis, the research seeks to identify recurring patterns of misunderstanding, explain their linguistic and cultural sources, and highlight the limitations of current model design.

2. METHODS

This study adopts a qualitative research design to investigate how natural language processing (NLP) models interpret and misinterpret non-standard Indonesian. The focus is on exploring errors and their underlying causes rather than measuring numerical performance. The methodology consists of three main stages: the development of a dialectal corpus, the application of selected models, and a systematic error analysis.

The foundation of this research is the NusaDialect (NusaD) corpus, which was created to represent the diversity of non-standard Indonesian. Texts were collected from publicly available platforms such as Twitter/X, Instagram, and online forums, where informal registers, dialectal expressions, and code-mixed utterances are widely used. A stratified sampling approach ensured balanced coverage of major dialect groups, including Jakartan Betawi, Javanese-influenced Indonesian, Sundanese-influenced Indonesian, and Eastern Indonesian varieties, as well as multiple types of code-mixing, such as Indonesian-English and Indonesian-Javanese. To comply with ethical research standards, only publicly accessible posts were included, and all user information was anonymized. The collected texts were manually annotated by native speakers representing different linguistic backgrounds. Annotators identified the dialectal influence of each text, marked instances of code-mixing, and provided task-specific labels for sentiment polarity and named entities. Disagreements were discussed collaboratively, leading to refinements in the annotation schema and ensuring the reliability of the corpus.

Once the dataset was prepared, a set of widely recognized NLP models was applied to generate outputs for analysis. These models included Multilingual BERT (Devlin et al., 2019), IndoBERT (Koto et al., 2020), and XLM-RoBERTa (Conneau et al., 2020), all of which represent established baselines in Indonesian NLP research. The models were used without extensive fine-tuning, as the goal was not to optimize performance but to expose their interpretive limitations when processing dialectal and code-mixed input. The predictions produced by these models, such as misclassified sentiments and misidentified named entities, formed the primary material for the qualitative analysis.

The central stage of the methodology consisted of a detailed qualitative error analysis. A purposive sample of errors was manually examined and categorized according to linguistic features, covering lexical, morphological, syntactic, semantic, and cultural dimensions, as well as issues arising from code-mixing and named entity recognition. This taxonomy of errors, informed by earlier frameworks in error analysis and bias in NLP (Blodgett et al., 2020; Winata et al., 2019), was refined throughout the analysis to capture the nuances of dialectal variation in Indonesian. The emphasis was placed on interpretive depth, with the objective of explaining why particular errors occurred and what they reveal about the broader challenges of modeling Indonesian linguistic diversity. Instead of relying on numerical accuracy scores, the analysis highlighted thematic patterns and recurring error sources, offering insights into the interaction between language models and non-standard varieties of Indonesian.

3. FINDINGS AND DISCUSSION

Finding

The findings from this study highlight both the scale of challenges and the opportunities in building AI systems for Indonesian non-standard dialects. We begin by presenting examples from the proposed NusaDialect benchmark to illustrate the linguistic diversity of the data, followed by quantitative evaluation, qualitative error analysis, and mitigation strategies.

The NusaDialect benchmark captures a broad range of Indonesian non-standard language varieties, including slang, regional dialects, and code-mixed expressions. These samples reflect the complexity that AI models must process when dealing with real-world digital communication.

Table 1 presents selected entries from the dataset, showcasing examples of dialectal variation, code-switching with English and Javanese, as well as task-specific annotations for sentiment and named entities.

ID	Text (Non-Standard)	Dialect / Style	Code-Mix	Sentiment (Gold)	NER (Gold)
1	"Gue baru balik dari **Bandung**. **Ciwalk** itu **recommended** banget buat **hangout**."	Jakartan / Betawi	EN	Positive	LOC (Bandung, Ciwalk)
2	"**Keren** banget **event JavaJazz** tahun ini, penampilan **Tulus** *mantap*!"	General Slang	EN	Positive	PER (Tulus), MISC (JavaJazz)
3	"**Kopi** di **Kedai ini overpriced**. Rasanya biasa aja, **ga worth it**."	Informal	EN	Negative	ORG (Kedai ini)
4	"**Aplikasi GoCar lagi error**, **driver**-nya **cancel** terus. **Bete**!"	Jakartan	EN	Negative	ORG (GoCar)
5	"**Wisata Dieng** itu **pesonanya** luar biasa. **Udaranya adem ayem**."	Javanese-influenced	JW	Positive	LOC (Dieng)
6	"**Pengen kulineran** di **Surabaya**, tapi **duit** lagi tipis."	Javanese-influenced	JW	Neutral	LOC (Surabaya)
7	"**Pelayanan RS Siloam very professional**. **Nurses**-nya ramah."	Formal Mix	EN	Positive	ORG (RS Siloam)
8	"**Gue deactivate IG** dulu, **need break** dari **social media**."	Jakartan	EN	Neutral	MISC (IG)
9	"**Motor Honda Vario** kena **curi** di depan **mall**."	Informal	-	Negative	ORG (Honda), PROD (Vario)
10	"**Cewek** itu **style**-nya **always on point**. **Fashionista** banget."	Slang	EN	Positive	-
11	"**Makan sambal** pedes **nanging** nikmat."	Javanese	JW	Positive	-
12	"**Booking tiket via Traveloka** prosesnya gampang banget."	Informal	EN	Positive	ORG (Traveloka)
13	"**Karya** seniman **Bali Ari Astina** bikin **speechless**."	Balinese-influenced	EN	Positive	PER (Ari Astina)

14	"**Hape Xiaomi** ku lagi **ngadat**. **Baterai** bocor parah."	Eastern (Manado)	-	Negative	ORG (Xiaomi), PROD (Hape)
15	"**Pengumuman resmi** dari **Kemenkes** soal **booster**."	Formal	EN	Neutral	ORG (Kemenkes)

Table 1 presents a synthetic sample from the NusaDialect Benchmark dataset, designed to illustrate the diversity of linguistic phenomena encountered in Indonesian digital communication. The examples reflect the interplay between regional dialects, informal slang, and code-mixing practices that pose challenges for NLP systems. Each entry is annotated with dialectal or stylistic labels, code-mixing sources, and gold labels for sentiment analysis and named entity recognition (NER), offering a structured basis for evaluating model performance.

The dataset captures the spectrum of Indonesian linguistic variation, ranging from Jakartan/Betawi colloquialisms such as "*gue*" and "*bete*" (IDs 1, 4, and 8) to Javanese-influenced phrases like "*adem ayem*" and "*pengen kulineran*" (IDs 5 and 6). It also highlights how Eastern Indonesian varieties, such as Manado-influenced speech (ID 14), contribute distinct vocabulary ("*ngadat*"), while Balinese-influenced expressions appear in artistic contexts (ID 13). Such examples illustrate how regional identities surface online, enriching but also complicating automated language processing.

Another critical feature of the dataset is code-mixing, particularly between Indonesian and English. Borrowed terms such as "recommended," "hangout," "cancel," and "always on point" (IDs 1, 4, and 10) are seamlessly embedded within Indonesian sentences. While this mirrors natural communication among younger internet users, it creates ambiguity for models trained primarily on monolingual corpora. Similarly, brand names, applications, and event titles—such as "GoCar," "Traveloka," and "JavaJazz" (IDs 2, 4, and 12)—are tagged under NER tasks, demonstrating the importance of handling named entities that frequently appear in mixed linguistic contexts.

From a task perspective, the dataset integrates sentiment polarity and NER annotation, enabling multifaceted evaluation. Positive sentiments are often tied to tourism and cultural pride (IDs 5 and 13), while negative sentiments emerge in service complaints or product dissatisfaction (IDs 3, 4, and 14). Neutral entries typically reflect situational statements without strong affect (IDs 6, 8, and 15). For NER, the benchmark includes a wide range of entities—locations (Bandung, Dieng, Surabaya), organizations (GoCar, Rumah Sakit Siloam), and persons (Tulus, Ari Astina)—ensuring comprehensive testing of entity recognition capabilities.

Overall, Table 1 illustrates the linguistic richness and challenges of Indonesian dialectal data. It demonstrates how everyday communication online blend regional identity, slang innovation, and global influences, all of which demand dialect-aware NLP solutions. By foregrounding these complexities, the NusaDialect dataset underscores the need for models capable of robustly handling non-standard, socially embedded language use in Indonesia.

Quantitative Results

To assess the robustness of current models, we compared performance across multilingual and Indonesian-specific architectures. Table 2 reports results on the sentiment analysis task, while Table 3 summarizes performance on named entity recognition (NER).

Table 2. Sentiment Analysis Results

Model	Accuracy	Precision	Recall	F1	Gap vs. Formal Indo.
mBERT	65.2%	64.8%	61.1%	62.1%	-25.4
IndoBERT	76.1%	75.9%	74.2%	75.4%	-18.4
XLM-R	79.5%	78.1%	78.9%	78.9%	-10.3
GPT-4 (5-shot)	**83.2%**	**82.5%**	**83.1%**	**83.5%**	N/A

Table 3. NER Results (F1 by Entity Type)

Model	PER	LOC	ORG	MISC	Overall F1
mBERT	55.1	68.3	48.2	30.5	55.2
IndoBERT	70.5	75.8	65.4	52.1	68.9
XLM-R	**72.8**	**78.9**	**68.1**	**58.7**	**71.6**

Table 2 reports the performance of several NLP models on the sentiment analysis task using the NusaDialect dataset. The results reveal a clear performance gap between formal Indonesian benchmarks and dialectal data. For example, mBERT achieves only 62.1% F1-score, reflecting difficulties in processing informal and code-mixed text, compared to its substantially higher accuracy on standard Indonesian. In contrast, IndoBERT, which is tailored for Indonesian, performs better (75.4% F1-score), but still shows a notable decline of nearly 18 points when tested on dialectal input.

Interestingly, XLM-RoBERTa outperforms IndoBERT with an F1-score of 78.9%, despite being trained as a general multilingual model. This suggests that exposure to diverse languages during training may enhance its adaptability to Indonesian dialectal variations. The strongest performance is observed with GPT-4 (83.5% F1-score) under a few-shot learning setup, demonstrating that large-scale pretrained models are inherently more resilient to linguistic irregularities, likely due to their broader lexical coverage and contextual learning capabilities. Overall, Table 2 highlights the limitations of conventional Indonesian-specific models when confronted with informal and dialectal data, while also pointing to the promise of multilingual and large language models in bridging this gap.

Table 3 presents the Named Entity Recognition (NER) results across four entity types: Person (PER), Location (LOC), Organization (ORG), and Miscellaneous (MISC). Performance varies significantly across categories, with LOC entities showing relatively high F1-scores across all models, as place names tend to be stable across dialects. Conversely, MISC entities perform the worst, with F1-scores as low as 30.5% for mBERT, reflecting the difficulty of recognizing event names, slang-based identifiers, and borrowed English expressions such as "JavaJazz" or "IG."

Among the models, IndoBERT again demonstrates strong performance (68.9% overall F1), but it is slightly surpassed by XLM-RoBERTa (71.6% overall F1), which performs better across PER, LOC, and ORG entities. This supports the notion that multilingual pretraining provides

broader generalization, especially in handling names and terms not commonly found in standard Indonesian corpora. The weakest results are consistently produced by mBERT, which struggles to adapt to dialectal and code-mixed input.

Taken together, Table 3 underscores that NER is particularly sensitive to code-mixing and slang innovation. Errors often stem from unfamiliar orthography, unconventional capitalization, and cultural references embedded in dialectal usage. These findings emphasize that while sentiment analysis suffers from polarity misinterpretation, NER faces deeper structural challenges in recognizing the boundaries and categories of entities in non-standard Indonesian.

The results show a clear performance gap: all models perform significantly worse on dialectal data than on formal Indonesian benchmarks. Sentiment analysis remains relatively stable, especially for GPT-4, which leverages large-scale pretraining to achieve strong few-shot results. However, NER performance is particularly weak, especially for organization and miscellaneous entities often expressed in English or hybrid forms. Interestingly, IndoBERT, though strong on formal Indonesian, is outperformed by XLM-RoBERTa on dialectal text, suggesting that broader multilingual exposure increases robustness.

Qualitative Error Analysis

Beyond numeric scores, we examined 100 misclassifications to identify patterns of model failure. Table 4 categorizes these errors into four groups: out-of-vocabulary (OOV) words, code-mixing ambiguity, morphological complexity, and pragmatic or cultural context.

Table 4. Error Categories

Category	Frequency	Example	Challenge
OOV Words	35%	"Bete", "Mantul"	Lacking embeddings for slang/dialect
Code-Mixing	30%	"Cancel" misread	Context of English words lost
Morphological	20%	"Kulineran", "Ngadat"	Agglutinative forms not parsed
Pragmatic/Cultural	15%	"Adem ayem", sarcasm	Lacks cultural grounding

Table 4 categorizes the most frequent sources of errors when models are applied to non-standard Indonesian. The largest portion of errors (35%) stems from Out-of-Vocabulary (OOV) words, particularly slang terms and dialectal expressions such as *bete* (annoyed) or *mantul* (great). Because these words are rarely included in pretraining corpora, models fail to generate accurate embeddings and instead rely on unrelated lexical cues, often leading to misclassification.

The second most frequent error type (30%) arises from code-mixing ambiguity, where models misinterpret English words integrated into Indonesian syntax. For example, the word *cancel* can indicate both a negative action (a driver canceling a ride) or a neutral/positive one (canceling a subscription), and models often fail to resolve these distinctions without strong contextual understanding. Morphological complexity contributes to 20% of errors, reflecting

challenges in segmenting agglutinative forms like *kulineran* (to go on a food adventure) or *ngadat* (acting up). Finally, pragmatic and cultural context errors account for 15% of failures, such as missing sarcasm or misinterpreting culturally rooted idioms like *adem ayem* (peaceful).

This taxonomy demonstrates that model failures are not merely technical but deeply sociolinguistic in nature. They arise from gaps in the models' exposure to Indonesian dialectal diversity, limited handling of morphological variation, and lack of cultural grounding in pragmatic usage. Understanding these categories provides a framework for designing targeted interventions, such as slang lexicon integration, dialect-sensitive tokenization, and culturally informed pretraining.

Mitigation Strategies

To address these weaknesses, two simulated strategies were tested: (a) continued pretraining of IndoBERT on social media text (IndoBERT-Social), and (b) data augmentation with back-translation (IndoBERT-Augment). Results are presented in Table 5.

Table 5. Mitigation Results (F1)

Model	Sentiment	NER	Formal Indo.
IndoBERT	75.4	68.9	93.8
IndoBERT-Social	**82.1** (+6.7)	**75.2** (+6.3)	92.5 (-1.3)
IndoBERT-Augment	79.8 (+4.4)	72.1 (+3.2)	**93.6** (-0.2)

Table 5 evaluates two mitigation strategies designed to address dialectal performance gaps: (1) Continued Pre-training on Social Media Data (IndoBERT-Social) and (2) Data Augmentation with Back-Translation (IndoBERT-Augment). Results indicate that both strategies improve performance on the NusaDialect dataset, though with different trade-offs. IndoBERT-Social achieves the highest gains (+6.7 F1 in sentiment, +6.3 F1 in NER), but its performance on formal Indonesian drops slightly, reflecting the risk of catastrophic forgetting when adapting models too heavily toward informal registers.

IndoBERT-Augment, on the other hand, delivers more balanced results. It provides moderate improvements on dialectal tasks (+4.4 F1 in sentiment, +3.2 in NER) while preserving nearly all performance on formal Indonesian. This suggests that augmentation strategies are more robust for production environments, where models must flexibly switch between formal and informal input. In practical terms, this means organizations deploying NLP systems in Indonesia—such as for customer service chatbots or social media monitoring—may prefer augmentation-based solutions to ensure stable performance across linguistic registers.

Overall, Table 5 highlights that while continued pretraining offers greater specialization, augmentation achieves a more sustainable trade-off between adaptability and stability. This aligns with broader findings in multilingual NLP, where hybrid strategies combining pretraining with targeted augmentation often yield the most resilient models in real-world use.

Both strategies improve performance on dialectal text, but with trade-offs. Continued pretraining (Strategy A) yields the strongest gains but slightly reduces accuracy on formal Indonesian, suggesting a risk of domain overfitting. In contrast, data augmentation (Strategy B) offers more balanced improvements, strengthening robustness on dialectal inputs while

maintaining performance on standard benchmarks. For practical applications requiring both formality levels, augmentation appears to be the safer long-term solution.

The results confirm that current NLP models for Indonesian struggle with the reality of dialectal variation and code-mixing. While larger models like GPT-4 demonstrate promise, consistent robustness requires targeted strategies such as augmentation or domain-adaptive training. These findings underscore the urgency of designing benchmarks like NusaDialect and pursuing solutions that account for Indonesia's true linguistic landscape.

Discussion

The performance gap observed across models on the NusaDialect benchmark underscores the persistence of linguistic bias in NLP systems. According to Bender & Friedman (2018)"data statements for NLP," model outcomes reflect the linguistic assumptions of their training data. Since most Indonesian NLP resources privilege the standardized, formal register, dialectal and informal usages are marginalized. This explains why slang expressions like *bete* or morphologically complex forms like *kulineran* are misclassified: the models are simply not exposed to them during pretraining. Our error analysis thus provides empirical confirmation of language ideologies in NLP—where certain varieties are "legitimate" for computation, while others are invisibles (Blodgett et al., 2020).

A second theoretical lens is code-switching theory, particularly Myers-Scotton (1997) Matrix Language Frame (MLF) model. This framework explains how bilingual speakers structure sentences around a dominant "matrix" language, embedding words or phrases from another. In Indonesian digital discourse, Standard or colloquial Indonesian often serves as the matrix, while English contributes embedded lexicon (cancel, recommended, hangout). The difficulty faced by IndoBERT and mBERT in parsing such sentences stems from their inability to recognize how functional morphemes and semantic roles remain in the Indonesian matrix while content words are borrowed from English (Makrifah & Fauzi, 2024; Simanjuntak et al., 2025; Sormin et al., 2025). NLP struggles here because current tokenizers segment at surface level, without access to structural insights predicted by MLF theory (Çetinoğlu et al., 2016).

The morphological complexity observed, particularly with agglutinative and reduplicative processes in Indonesian dialects, further reflects theoretical gaps. Indonesian and regional varieties employ productive affixation (e.g., *kuliner* → *kulineran* "to go on a food adventure") and cliticization (*nya, ku*). From the perspective of morphological typology (Haspelmath & Sims, 2010), these are predictable, rule-governed transformations. However, sub-word tokenization methods like BPE and Word Piece often fragment such forms in ways that break semantic cohesion. The high rate of segmentation errors in our taxonomy aligns with linguistic theory that agglutinative languages require morpheme-aware or character-level modeling rather than purely statistical sub-word segmentation.

Another salient issue is pragmatic and cultural context, which falls under the domain of pragmatics theory (Levinson, 1983). Sarcasm, idiomatic expressions, and culturally specific terms such as *adem ayem* (peaceful) are pragmatically loaded and cannot be decoded without contextual or cultural grounding. Large language models like GPT-4, though more resilient, also fail here because pretraining corpora underrepresent pragmatic and cultural nuance. This limitation parallels (Gumperz, 1982) sociolinguistic insights on contextualization cues, where meaning emerges from shared cultural knowledge rather than lexical items alone.

Finally, the mitigation experiments connect to domain adaptation theory in NLP. Continued pretraining on dialectal corpora (IndoBERT-Social) illustrates the trade-off described by Gururangan et al. (2020): gains on the target domain may come at the cost of "catastrophic forgetting" in the source domain. By contrast, our data augmentation strategy resonates with theory-driven synthetic data generation (Munawir et al., 2024; Pratapa et al., 2018; Rahmat et al., 2025; Sukabdi et al., 2025), which argues that expanding linguistic coverage without overwriting prior distributions is more robust. This balance is particularly critical in Indonesia, where models must navigate a continuum from formal state communication to informal digital slang.

Taken together, these findings show that the challenges are not simply engineering issues but are deeply rooted in linguistic structure and sociolinguistic reality. Theories of code-switching, morphology, and pragmatics provide explanatory depth to model errors, while domain adaptation theory clarifies the trade-offs of mitigation. Addressing these gaps will require not only technical refinements but also a shift toward linguistically informed modeling that respects the multilingual, code-mixed, and culturally grounded nature of Indonesian communication.

4. CONCLUSION

This study confirms that current NLP models trained on standard Indonesian struggle significantly when confronted with dialectal variation, slang, and code-mixed digital communication. Both quantitative and qualitative findings revealed systematic weaknesses in handling out-of-vocabulary terms, code-switching ambiguity, morphological complexity, and culturally embedded pragmatic expressions. Although mitigation strategies such as domain-specific pretraining and data augmentation proved effective, trade-offs remain between specialization in dialectal data and preserving performance on formal Indonesian. A key limitation of this research is its reliance on synthetic data, which, while useful for controlled experimentation, cannot fully represent the richness of Indonesia's linguistic landscape. Moreover, the study focused only on sentiment analysis and NER, leaving many other NLP tasks unexplored.

For future research, the construction of authentic, community-sourced corpora is essential to advance inclusive Indonesian NLP, complemented by methods that integrate linguistic theory into model design, such as morphology-aware tokenization or culturally informed pragmatics. Beyond academic implications, this study also carries practical significance. More dialectally robust models can improve user experience in real-world applications such as chatbots, digital government services, and social media monitoring, ensuring equitable access for speakers of diverse Indonesian varieties. By addressing these gaps, NLP technologies can better reflect Indonesia's linguistic diversity while supporting industries, policymakers, and communities in building more inclusive and culturally sensitive AI systems.

5. REFERENCES

Abdalla, M., Wahle, J. P., Ruas, T., Névéol, A., Ducel, F., Mohammad, S. M., & Fort, K. (2023). The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13141–13160. https://doi.org/10.18653/v1/2023.acl-long.734

Arif, M., Aziz, M. K. N. A., & Arif, M. A. M. (2025). A Recent Study on Islamic Religious Education Teachers' Competencies in the Digital Age: A Systematic Literature Review. *Journal of Education and Learning (EduLearn)*, *19*(2), 587–596.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M. S., Parmonangan, I., Alfina, I., Putra, I. F., Rahmadani, S., … Purwarianti, A. (2023). NusaCrowd: Open Source Initiative for Indonesian NLP Resources. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13745–13818). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.868

Çetinoğlu, Ö., Schulz, S., & Vu, N. T. (2016). *Challenges of Computational Processing of Code-Switching* (No. arXiv:1610.02213). arXiv. https://doi.org/10.48550/arXiv.1610.02213

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Elice, D., Patimah, S., Pahrudin, A., Koderi, Fauzan, A., & Liriwati, F. Y. (2025). Development of Abacus Training Management in the Artificial Intelligence Era. *Munaddhomah: Jurnal Manajemen Pendidikan Islam*, *6*(2), 267–280. https://doi.org/10.31538/munaddhomah.v6i2.1719

García, O., & Wei, L. (2014). *Translanguaging*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137385765

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Haspelmath, M., & Sims, A. D. (2010). *Understanding Morphology*. ResearchGate. https://www.researchgate.net/publication/333317956_Understanding_Morphology_2ed

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2021). *The State and Fate of Linguistic Diversity and Inclusion in the NLP World* (No. arXiv:2004.09095). arXiv. https://doi.org/10.48550/arXiv.2004.09095

Khotimah, S. H., Krisnawati, N. M., Abusiri, A., Mubin, F., & Wardi, M. (2024). Development of Virtual Field Trip-Based Learning Model as A Strengthening of Madrasah Student

Digital Literacy. *Nazhruna: Jurnal Pendidikan Islam*, *7*(1), 103–121. https://doi.org/10.31538/nzh.v7i1.4532

Kohler, M. (2019). Language education policy in Indonesia: A struggle for unity in diversity. In *The Routledge International Handbook of Language Education Policy in Asia*. Routledge.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Kurniawan, S., Herlambang, S., Sari, N., Fadian, F., Suratman, B., Nurhidayah, V. A., Naffati, A. K., & Torikoh. (2024). Making Peace with Change: The Effect of GPT Chat Utilization on the Performance of Islamic Religion Teachers in Creating Teaching Modules. *Jurnal Pendidikan Agama Islam*, *21*(2), 492–509. https://doi.org/10.14421/jpai.v21i2.9767

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Makrifah, N., & Fauzi, N. (2024). Implementation of Talking Stick Learning Model to Improve English Learning Outcomes in Islamic Elementary School. *Fikroh: Jurnal Pemikiran Dan Pendidikan Islam*, *17*(1), 29–39. https://doi.org/10.37812/fikroh.v17i1.1403

Munawir, M., Alfiana, F., & Pambayun, S. P. (2024). Menyongsong Masa Depan: Transformasi Karakter Siswa Generasi Alpha Melalui Pendidikan Islam yang Berbasis Al-Qur'an. *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *7*(1), 1–11. https://doi.org/10.54069/attadrib.v7i1.628

Myers-Scotton, C. (1997). *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Nurhalisa, N., Rizal, R., Aqil, M., Lagandesa, Y. R., & Fasli, M. (2025). Pengaruh Model Problem Based Learning (PBL) dengan berbantuan Media Wordwall terhadap Hasil Belajar Siswa pada Mata Pelajaran Bahasa Indonesia. *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *8*(1), 151–159. https://doi.org/10.54069/attadrib.v8i1.867

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., & Bali, K. (2018). Language Modeling for Code-Mixing: The Role of Linguistic Theory based Synthetic Data. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1543–1553). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1143

Rahmat, M., Supriadi, U., Fakhruddin, A., Surahman, C., Abdillah, H. T., & Nurjanah, N. (2025). Religiosity and Interfaith Tolerance Among Students in Indonesian Islamic and General Junior Secondary Schools. *Jurnal Pendidikan Islam*, *11*(1), 115–132. https://doi.org/10.15575/jpi.v11i1.44660

Reksiana, Nata, A., Rosyada, D., Rahiem, M. D. H., & Ugli, A. R. R. (2024). Digital Extension of Digital Literacy Competence for Islamic Religious Education Teachers in the Era of Digital Learning. *Jurnal Pendidikan Agama Islam*, *21*(2), 402–420. https://doi.org/10.14421/jpai.v21i2.9719

Rohmiati, E. (2025). The Use of Digital Media in Learning Islamic Religious Education: Opportunities and Challenges. *Urwatul Wutsqo: Jurnal Studi Kependidikan Dan Keislaman*, *14*(1), 33–45. https://doi.org/10.54437/urwatulwutsqo.v14i1.1952

Sain, Z. H., Serban, R., Abdullah, N. B., & Thelma, C. C. (2025). Benefits and Drawbacks of Leveraging ChatGPT to Enhance Writing Skills in Secondary Education. *At-Tadzkir: Islamic Education Journal*, *4*(1), 40–52. https://doi.org/10.59373/attadzkir.v4i1.79

Simanjuntak, M. B., Rafli, Z., & Utami, S. R. (2025). Elevating Vocational Student Competence: The Crucial Need for English Literacy Competence. *Jurnal Ilmiah Peuradeun*, *13*(1), 721–744. https://doi.org/10.26811/peuradeun.v13i1.1109

Sodikin, S. (2024). Transformasi Pendidikan Agama Islam Melalui Artificial Intelligent (AI): Upaya Meningkatkan Kemampuan Berpikir Kritis Mahasiswa. *Academicus: Journal of Teaching and Learning*, *3*(2), 78–89. https://doi.org/10.59373/academicus.v3i2.65

Sormin, D., Siregar, I., Rambe, N., Siregar, R., Lubis, J. N., & Kholijah, A. (2025). Implementation of the Ismubaris Curriculum (Islamic Studies, Muhammadiyah Ideology, Arabic, and English). *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *8*(2), 464–473. https://doi.org/10.54069/attadrib.v8i2.920

Sukabdi, Z. A., Sofanudin, A., Munajat, M., Mulyana, M., & Budiyanto, S. (2025). The Challenge of Terrorism Regeneration: What Schools Do Terrorist Offenders Select for Their Children? *Ulumuna*, *29*(1), 102–128. https://doi.org/10.20414/ujis.v29i1.1061

Syukur, F., Maghfurin, A., Marhamah, U., & Jehwae, P. (2024). Integration of Artificial Intelligence in Islamic Higher Education: Comparative Responses between Indonesia and Thailand. *Nazhruna: Jurnal Pendidikan Islam*, *7*(3), 531–553. https://doi.org/10.31538/nzh.v7i3.13

Topuha, O. K., Rizal, R., Aqil, M., Gagaramusu, Y. B. M., & Fasli, M. (2025). Pengaruh Penggunaan Media Digital Terhadap Hasil Belajar Siswa Pada Mata Pelajaran Bahasa Indonesia di Sekolah Dasar. *Attadrib: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *8*(1), 174–183. https://doi.org/10.54069/attadrib.v8i1.866

Winata, G. I., Lin, Z., & Fung, P. (2019). Learning Multilingual Meta-Embeddings for Code-Switching Named Entity Recognition. In I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, & M. Rei (Eds.), *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 181–186). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4320

Wulandari, F., Widyaningrum, N., Sa'ida, N., & Masturoh, U. (2025). Meningkatkan Kemampuan Bahasa Anak Usia Dini melalui Pembelajaran Multimedia Interaktif Berbasis AR dan VR. *Academicus: Journal of Teaching and Learning*, *4*(1), 61–70. https://doi.org/10.59373/academicus.v4i1.86

Zein, S. (2020). *Language policy in superdiverse Indonesia*. ResearchGate. https://www.researchgate.net/publication/340175378_Zein_-_2020_-_Language_policy_in_superdiverse_Indonesia_-_Chapter_1_copy