e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Volume 10, Issue 1, **60-69.**
https://e-journal.hamzanwadi.ac.id/index.php/edumatic
DOI: 10.29408/edumatic.v10i1.33724

Mitigating Class Imbalance in Indonesian Sarcasm Detection: A Cross-Platform Transformer Study

A. Salky Maulana [1,*] , I Made Artha Agastya [1]

[1] Universitas Amikom Yogyakarta, Indonesia
[*] Corresponding author: A. Salky Maulana, Universitas Amikom Yogyakarta, Indonesia
✉ maulanasalky@students.amikom.ac.id

Received: 25 December 2025 | Revised: 14 January 2026 | Accepted: 13 February 2026 | Published: 4 March 2026

Abstract
Sarcasm detection in Indonesian social media remains challenging due to implicit pragmatic expressions, severe class imbalance, and strong domain variation across platforms. Unlike prior Indonesian sarcasm studies that predominantly focus on in-domain accuracy using conventional balancing methods, this study provides the first systematic cross-platform analysis of generative data balancing under domain shift. We empirically examine whether GPT-4o based generative balancing improves robustness rather than accuracy-centric evaluation in Transformer-based sarcasm detection. Models trained on Twitter data are evaluated across Twitter, Reddit, and TikTok as an unseen domain. The results show that generative balancing yields limited gains in in-domain evaluation but consistently improves cross-domain robustness by increasing sarcasm recall, particularly for Base models. Notably, XLM-R Base achieves an absolute F1-score improvement of +10.8 points on TikTok, while IndoBERT-Large attains the highest in-domain F1-score of 0.7444. These findings indicate that generative augmentation partially mitigates class imbalance by enhancing robustness under domain shift, thereby repositioning sarcasm detection as a robustness-oriented problem and highlighting generative balancing as a complementary strategy rather than a substitute for larger Transformer models in cross-platform NLP settings.

Keywords: class imbalance; cross-domain robustness; generative data augmentation; sarcasm detection; transformer models

To cite this article: Maulana, A. S., & Agastya, I. M. A. (2026). Mitigating Class Imbalance in Indonesian Sarcasm Detection: A Cross-Platform Transformer Study. *Edumatic: Jurnal Pendidikan Informatika*, *10*(1), 60–69. https://doi.org/10.29408/edumatic.v10i1.33724

INTRODUCTION

The expansion of social media has intensified the complexity of text-based communication, introducing linguistic phenomena such as sarcasm that pose persistent challenges for Natural Language Processing (NLP) systems. Sarcasm operates as a pragmatic inference problem rather than a surface-level classification task, as it relies on implicit meaning that contradicts literal expressions. When sarcasm is not accurately detected, sentiment polarity is systematically inverted, undermining the reliability of downstream applications in noisy and heterogeneous environments (Helal et al., 2024; Liu et al., 2024; Šandor & Bagić Babac, 2024).

Recent advances in Transformer-based architectures such as BERT and XLM-R have substantially improved contextual language modeling and motivated their adoption as a widely used approach for sarcasm detection. However, strong representational capacity alone does not guarantee pragmatic understanding. Statistical probability modeling does not inherently

capture pragmatic reasoning, which depends on contextual incongruity and speaker intent (Herawan & Saputri, 2025). As a result, Transformers often rely on surface-level regularities rather than invariant pragmatic cues.

Gedela et al. (2024) and Sukhavasi and Dondeti (2023) report that Transformer-based sarcasm detectors rely on spurious lexical patterns rather than stable contextual cues, a behavior associated with shortcut learning under class imbalance and domain shift. Consequently, high in-domain accuracy becomes a weak proxy for real-world sarcasm competence, as performance often degrades sharply when models are evaluated outside their training distribution (Hupkes et al., 2023; Pandey & Singh, 2023). This highlights a gap between contextual encoding capacity and robust pragmatic generalization.

These challenges are further amplified in Indonesian social media, where sarcasm commonly involves informal slang, emphatic particles, and platform-specific discourse conventions (Nasution et al., 2025). Language variation across platforms such as Twitter, Reddit, and TikTok violates the Independent and Identically Distributed (I.I.D.) assumption, introducing structural distributional shifts that systematically undermine model transferability. Consequently, single-platform evaluation becomes insufficient for assessing generalization in Indonesian sarcasm detection (Liebeskind & Bączkowska, 2025). However, most Indonesian studies still report performance primarily using in-domain accuracy, implicitly treating in-domain performance as indicative of cross-platform reliability (An et al., 2024; Suhartono et al., 2024).

Class imbalance further compounds this problem by biasing learning toward the majority non-sarcasm class and producing accuracy gains that do not reflect genuine pragmatic recognition (A'la, 2025; Bayer et al., 2023; Hu et al., 2025; Thakkar et al., 2024). Under such conditions, model performance depends not only on architecture but also on the distribution of training data. IndoBERT captures language-specific linguistic patterns, whereas XLM-R offers broader lexical coverage, yet both remain sensitive to imbalance and domain shift. Recent studies indicate that generative data augmentation can introduce semantically coherent variation in training data (Sujana & Kao, 2023; Zhao et al., 2024). However, whether such augmentation improves cross-platform robustness rather than merely in-domain performance remains insufficiently examined in Indonesian sarcasm detection.

This study examines whether GPT-4o based generative balancing improves robustness under platform-induced distribution shift in Indonesian sarcasm detection. IndoBERT and XLM-R are trained on Twitter data and evaluated across Twitter, Reddit, and a manually annotated TikTok dataset. By reframing sarcasm detection as a robustness-oriented inference task, this work demonstrates that generative balancing functions as an implicit regularization mechanism that mitigates lexical shortcut learning, thereby contributing empirical and theoretical insight into robustness-oriented modeling for low-resource and pragmatically complex languages.

METHOD

This study employs a controlled experimental benchmarking design to examine whether generative text balancing improves the robustness of Transformer-based models for Indonesian sarcasm detection under class imbalance and domain shift. To ensure fair comparison, architectures, data splits, and hyperparameters were held constant. As illustrated in Figure 1, the experimental workflow spans data collection and exploratory analysis, generative augmentation applied exclusively to the training data, standardized preprocessing, model training under baseline and balanced scenarios, and evaluation on in-domain (Twitter) and cross-domain (Reddit and TikTok) test sets.

Training and validation data were obtained from the IdSarcasm benchmark (Suhartono et al., 2024) containing 1,878 training and 268 validation instances. Evaluation utilized three

test datasets: IdSarcasm Twitter, adapted Reddit, and a TikTok dataset scraped from trending videos. TikTok annotation was performed manually by two authors based on contextual incongruity, implicit negation, or sarcastic tone, with ambiguous cases conservatively labeled as non-sarcasm. Exploratory Data Analysis (EDA) revealed substantial class imbalance across all domains. Table 1 details the specific dataset splits to ensure reproducibility, while Figure 2 visually quantifies this imbalance, highlighting sarcasm as the minority class across all platforms.

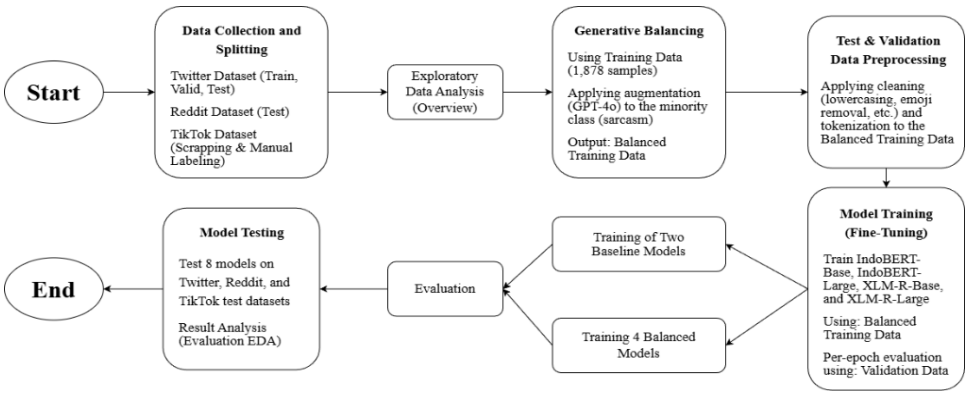

Figure 1. Sarcasm detection workflow with generative balancing

Table 1. Dataset statistic and split

Source	Data Split	Total Samples	Sarcasm Class	Non-Sarcasm Class
Twitter	Train	1878	470	1408
Twitter	Validation	268	67	201
Twitter	Test	538	134	404
Reddit	Test	500	114	386
Tiktok	Test	500	219	281

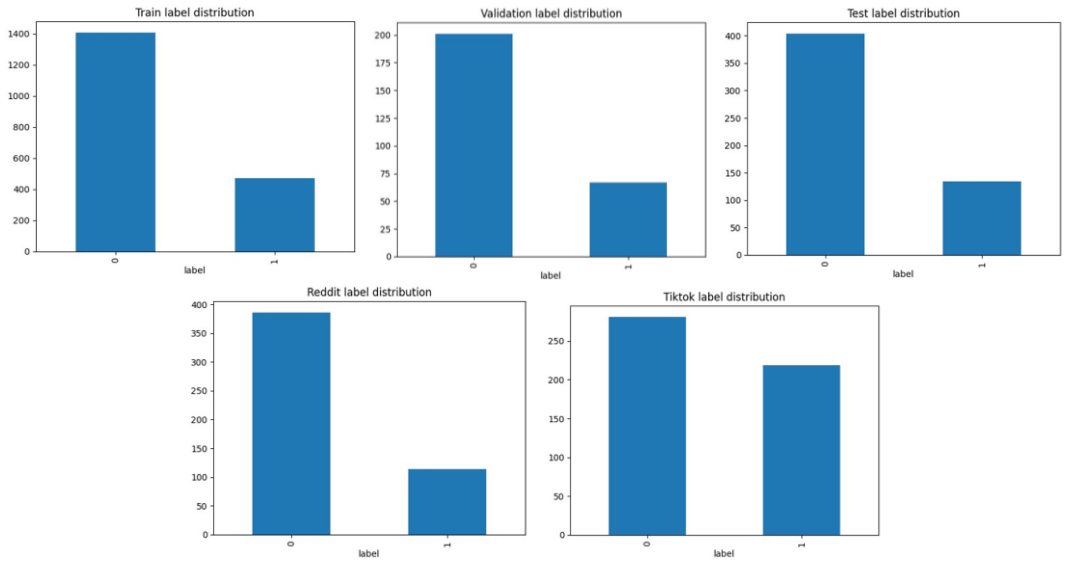

Figure 2. Class distribution

To mitigate class imbalance, sarcastic samples were generated to achieve class parity (1:1) using GPT-4o, employing the following structured prompt: "*Kamu adalah pengguna Twitter Indonesia yang suka menulis dengan gaya santai, lucu, dan kadang sarkasme. Berikut beberapa contoh cuitan dari dataset asli… [Input Examples]. Sekarang buat [N] cuitan baru*

yang: (1) alami dan mirip gaya penulisan di atas, (2) mengandung nada sarkasme, (3) menggunakan bahasa santai, (4) maksimal 20 kata, dan (5) menghindari pengulangan frasa persis." Only training data were augmented to prevent leakage, and all generated samples were manually reviewed to ensure quality. Representative examples are shown in Table 2.

Table 2. GPT-4o augmentation examples

Original Data Example (Label 1)	Synthetic Data Generated by GPT
"*Selamat datang di Indonesia. Negara +62 dengan bermacam-macam adat istiadat keindahan alam budaya juga.. Kegoblokan sebagian masyarakatnya*"	"*Buku sejarah masa depan: "Indonesia, negeri dengan janji-janji manis*"

Text preprocessing was applied uniformly, including dataset-specific token removal, cautious slang normalization, stopword elimination, and case folding. Tokenization used the Hugging Face AutoTokenizer (max length 128). Four models were evaluated: IndoBERT (Base/Large) and XLM-R (Base/Large), fine-tuned using the Hugging Face library on NVIDIA T4 and P100 GPUs with hyperparameters detailed in Table 3. Models were evaluated under zero-shot cross-domain conditions. Performance was measured using Accuracy, Precision, Recall, and Macro-F1 Score, prioritizing minority sarcasm detection and robustness.

Table 3. Transformer model configurations and training parameters

Model	GPU	LR	Batch	Epoch	Early Stop
IndoBERT-Base	T4	2e-5	16	12	4
IndoBERT-Large	P100	1e-5	4 (grad ×2)	10	5
XLM-R-Base	T4	2e-5	16	12	4
XLM-R-Large	P100	1.5e-5	4 (grad ×2)	15	6

RESULTS AND DISCUSSION

Results

This section reports the benchmarking results of Transformer-based models for Indonesian sarcasm detection, specifically addressing the core research question: does generative balancing improve fundamental robustness or merely inflate in-domain accuracy? Rather than presenting raw scores alone, the analysis focuses on identifying systematic performance patterns across models, training scenarios, and evaluation domains. Three key findings emerge: (1) generative text balancing yields limited gains for in-domain evaluation but consistently improves cross-domain robustness, (2) Base models are more sensitive to class imbalance and benefit more substantially from generative balancing than Large models, and (3) increased model capacity alone does not guarantee better generalization, as architectural differences, particularly monolingual versus multilingual pre-training, play a decisive role under domain shift.

The validation results are reported exclusively for model selection and training stability, not as conclusive evidence of real-world generalization. All models were evaluated on the IdSarcasm validation set to select the best-performing checkpoint based on Macro-F1 Score. Training hyperparameters and validation results are summarized in Table 4 and Table 5, respectively.

Training configurations in Table 4 were selected to ensure stable convergence. Base models utilized a batch size of 16, while Large models required a batch size of 8. For Large variants, Version 2 (V2), characterized by a higher learning rate and extended training duration, was used due to improved training stability. These hyperparameters were held constant across

baseline and balanced scenarios to isolate the impact of the data balancing strategy from architectural variations.

Table 4. Transformer model training configuration

Model	Scenario	Learning Rate	Batch Size	Epoch
IndoBERT-Base	Baseline	1e-5	16	12
XLM-R-Base	Baseline	1e-5	16	12
IndoBERT-Base	Balanced	1e-5	16	12
XLM-R-Base	Balanced	1e-5	16	12
IndoBERT-Large V1	Balanced	1e-5	8	10
XLM-R-Large V1	Balanced	1e-5	8	10
IndoBERT-Large V2	Balanced	1.5e-5	8	15
XLM-R-Large V2	Balanced	1.5e-5	8	15

Table 5. Comparison of F1-scores on validation data

Model	Scenario	F1-score (validation)	Accuracy
IndoBERT-Base	Baseline	0.7172	0.8470
XLM-R-Base	Baseline	0.5437	0.8246
IndoBERT-Base	Balanced	0.6714	0.8284
XLM-R-Base	Balanced	0.6870	0.8470
IndoBERT-Large V1	Balanced	0.7878	0.8507
XLM-R-Large V1	Balanced	0.8276	0.8657
IndoBERT-Large V2	Balanced	0.8255	0.8731
XLM-R-Large V2	Balanced	0.8397	0.8769

Table 5 highlights the impact of these configurations on model convergence and selection. The results show that generative balancing significantly aids the training stability of weaker models, specifically XLM-R Base, while Large models demonstrate consistently high performance. XLM-R Large Version 2 achieves the highest validation F1-score (0.8397), followed by IndoBERT Large Version 2 (0.8255). While these scores do not reflect real-world generalization, they function as a necessary control mechanism to ensure that subsequent performance shifts are driven by data augmentation rather than training instability. Generalization performance is assessed using held-out test datasets from Twitter (in-domain), Reddit, and TikTok (cross-domain). The comparison of F1-scores across models and training scenarios is summarized in Table 6.

Table 6. Comparison of F1-scores on test data

Model	Scenario	Twitter	Reddit	TikTok
IndoBERT-Base	Baseline	0.6980	0.3351	0.1592
XLM-R-Base	Baseline	0.5933	0.2166	0.1298
IndoBERT-Base	Balanced	0.6735	0.2975	0.2500
XLM-R-Base	Balanced	0.6769	0.3135	0.2377
IndoBERT-Large V1	Balanced	0.7072	0.2921	0.2384
XLM-R-Large V1	Balanced	0.7004	0.2902	0.2463
IndoBERT-Large V2	Balanced	0.7444	0.2865	0.2507
XLM-R-Large V2	Balanced	0.6990	0.3192	0.2747

Analysis of Table 6 reveals three systematic patterns governing model behavior. First, in-domain stagnation: generative balancing yields negligible gains on Twitter, where training

and test distributions align (e.g., IndoBERT Base decreases slightly from 0.698 to 0.6735). Second, cross-domain leaps: the same models demonstrate consistent and significant improvements on Reddit and TikTok (e.g., XLM-R Base gains +0.1079 on TikTok). Third, recall-driven improvement: these gains stem primarily from increased sensitivity to the minority sarcasm class rather than overall accuracy inflation. Furthermore, the results indicate an interaction effect between model capacity and imbalance; Base models rely heavily on augmentation to combat bias, whereas Large models leverage their capacity to mitigate imbalance effects. Notably, robustness to platform shifts appears more strongly influenced by pre-training diversity (XLM-R) than by language-specific modeling alone (IndoBERT), challenging the assumption that monolingual models are inherently superior for local tasks. To illustrate this robustness effect, Figure 3 compares baseline and balanced Base models across evaluation settings.

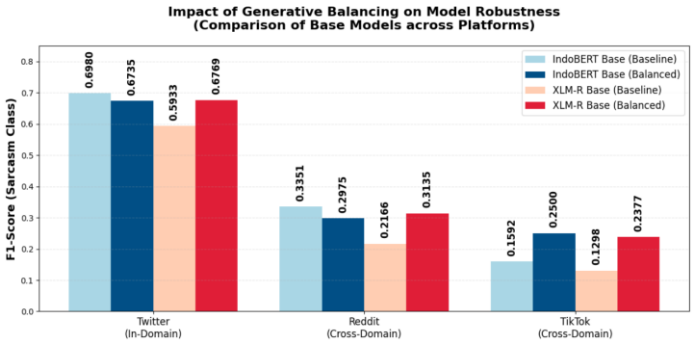

Figure 3. Generative balancing yields consistent cross-domain gains despite in-domain stagnation

Figure 3 compares the F1-score trends of balanced and baseline models across in-domain (Twitter) and cross-domain (Reddit and TikTok) test sets. The figure visualizes that generative balancing yields limited gains on Twitter but produces consistent improvements on Reddit and TikTok, confirming that its primary contribution lies in enhancing robustness under domain shift rather than optimizing single-domain accuracy. Among all configurations, XLM-R Large Version 2 exhibits the strongest cross-domain robustness. This effect cannot be attributed solely to increased model capacity, but is closely associated with its multilingual pre-training, which provides broader lexical coverage and improved handling of mixed-code expressions across platforms. To further characterize model error behavior under severe domain shift, detailed error analysis on TikTok is presented in Figure 4.

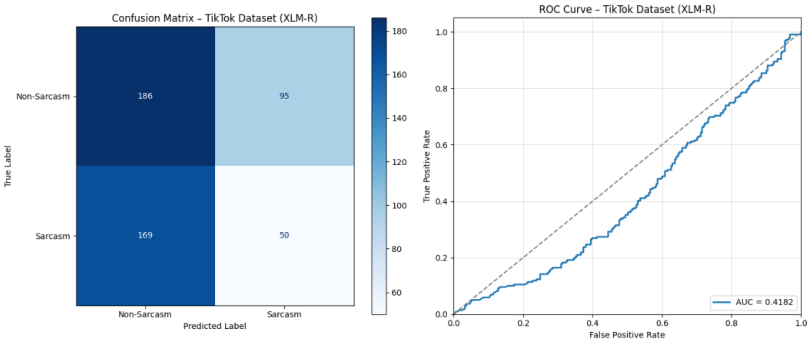

Figure 4. Error analysis on TikTok revealing high False Negatives (left) and limited discriminative separability (right)

The confusion matrix (left) highlights a high proportion of False Negatives, indicating that sarcastic comments are frequently misclassified as non-sarcasm. This behavior is

particularly pronounced on TikTok, where sarcasm often relies on video context, exaggerated pragmatics, and culturally embedded cues that are not fully captured in text-only models. Compared to the baseline model (not shown), generative balancing reduces the number of False Negatives, suggesting improved sensitivity to subtle sarcastic patterns.

The Receiver Operating Characteristic (ROC) curve (right) evaluates discriminative performance under severe class imbalance. For XLM-R Large Version 2 on the TikTok dataset yields an AUC value close to random (AUC $\approx$ 0.42), indicating limited global class separability. This result highlights that, despite improvements in recall, the model remains challenged in distinguishing sarcasm from non-sarcasm across decision thresholds in a highly divergent domain. Importantly, this low AUC does not invalidate the resilience claim; instead, it clarifies that generative balancing primarily improves sensitivity at the targeted operating point (minority class recall) rather than yielding uniformly strong class separation.

Performance degradation across platforms reflects increasing domain divergence: Twitter, as the training domain, yields the highest performance; Reddit introduces longer discourse and community-specific norms; and TikTok presents the greatest challenge due to short, slang-heavy, and highly contextual comments. This trend confirms that cross-platform sarcasm detection is fundamentally a robustness-oriented problem rather than a purely accuracy-driven task, reinforcing the need for evaluation beyond single-domain benchmarks.

Discussion

The findings of this study demonstrate that generative text balancing functions primarily as a robustness-enhancing mechanism rather than a simple accuracy booster. Across experiments, generative augmentation yields limited benefits under in-domain evaluation but consistently improves performance under cross-domain distribution shift. This pattern indicates that sarcasm detection should be conceptualized as a generalization problem grounded in representation stability, rather than as a performance-driven classification task optimized for single-dataset accuracy (Hupkes et al., 2023). In pragmatic language phenomena such as sarcasm, accuracy measured within a fixed domain constitutes a weak proxy for true semantic competence, whereas robustness reflects a model's ability to maintain interpretive consistency when surface cues and discourse norms change (Fei et al., 2022; Liu et al., 2024).

The observed performance patterns suggest that Transformer-based models trained on imbalanced data tend to rely on lexical shortcut learning, over-associating frequent surface patterns with majority-class labels (Dogra et al., 2024). Although Transformers exhibit strong contextual encoding capacity, this capacity does not inherently confer pragmatic understanding (Hu et al., 2025). Under distribution shift, such shortcut-based representations fail, leading to severe recall degradation for minority classes. Generative balancing mitigates this failure mode by expanding the semantic and stylistic variability of sarcastic instances, thereby broadening the decision boundary and discouraging brittle lexical memorization. In this sense, generative augmentation operates as an implicit regularization mechanism, improving robustness by reshaping the representation space rather than merely inflating accuracy through data duplication (Chen et al., 2023; Zhao et al., 2024).

The differential effects observed between Base and Large models further clarify that robustness is not merely a function of model capacity. Large models can achieve strong in-domain performance through memorization even under class imbalance, but this memorization bias does not guarantee transferability when pragmatic cues shift across platforms. Base models, by contrast, benefit more substantially from generative balancing because the added semantic diversity compensates for their limited capacity, promoting more generalizable representations. These findings indicate an interaction effect between model capacity, data imbalance, and domain shift, challenging the assumption that scaling model size alone leads to robust sarcasm detection (Bayer et al., 2023).

Differences between IndoBERT and XLM-R can be interpreted through pre-training theory rather than superficial model comparison. IndoBERT's monolingual pre-training supports fine-grained modeling of Indonesian linguistic norms but constrains exposure to heterogeneous subword patterns, mixed-code expressions, and stylistic variation prevalent on platforms such as TikTok. XLM-R benefits from multilingual pre-training that provides broader lexical coverage, richer subword diversity, and prior exposure to cross-lingual variation, facilitating transfer under platform shift. These results suggest that robustness in sarcasm detection is more strongly influenced by pre-training diversity and representational breadth than by language-specific modeling alone (An et al., 2024; Nasution et al., 2025).

In contrast to prior Indonesian sarcasm studies that report strong in-domain accuracy (Javid & Mashayekhi, 2025; Suhartono et al., 2024), the present findings demonstrate that such performance often collapses under cross-platform evaluation. This indicates that single-domain benchmarks systematically overestimate real-world generalization and obscure robustness failures. By evaluating across Twitter, Reddit, and TikTok, this study reframes sarcasm detection as a robustness-oriented task, highlighting the necessity of cross-domain evaluation for socially grounded and pragmatically complex language phenomena (Liebeskind & Bączkowska, 2025).

Error analysis further supports this interpretation. The confusion matrix reveals persistent False Negatives under severe domain shift, while the low ROC AUC reflects limited global class separability. Importantly, this does not contradict the observed robustness gains. AUC measures global discrimination across thresholds, whereas sarcasm under domain shift is inherently non-separable due to pragmatic ambiguity. The observed improvements therefore reflect enhanced sensitivity at targeted operating points, specifically minority-class recall, rather than uniform separability across decision boundaries.

Several limitations should be acknowledged. This study does not claim that generative balancing resolves sarcasm detection under all conditions, nor that it produces universally separable representations. The reliance on text-only data limits interpretability on multimodal platforms such as TikTok, and the absence of quantified inter-annotator agreement introduces potential subjectivity. Moreover, generative augmentation may introduce subtle stylistic bias. Future work should investigate multimodal sarcasm detection, pragmatics-aware evaluation metrics, and robustness measures beyond F1-score, including calibration and fairness, to further ground robustness-oriented modeling in real-world deployment contexts (Liu et al., 2024).

CONCLUSION

This study explicitly answers the research question by demonstrating that GPT-4o based generative data balancing improves robustness under distribution shift rather than accuracy within a single domain. Conceptually, the findings reframe Indonesian sarcasm detection as a robustness-oriented inference problem, showing that generative balancing functions as an implicit regularization mechanism that reduces lexical shortcut learning and stabilizes minority-class decision boundaries through increased semantic diversity. The results further indicate that robustness does not emerge from model capacity alone, but from an interaction between pre-training diversity, data imbalance, and domain shift, highlighting the role of multilingual exposure in supporting cross-platform generalization beyond language-specific modeling. These conclusions are epistemically constrained: the study does not claim universal separability or multimodal sarcasm understanding, as the analysis is limited to text-only inputs and subject to annotation uncertainty. Accordingly, future research should integrate multimodal cues on video-centric platforms, investigate dynamic slang adaptation, and adopt calibration-aware robustness metrics to better assess real-world reliability. Taken together, this work establishes generative balancing as a theoretically grounded, data-centric approach for

narrowing the gap between laboratory performance and practical deployment in low-resource NLP.

REFERENCES

A'la, F. Y. (2025). Optimasi Klasifikasi Sentimen Ulasan Game Berbahasa Indonesia: IndoBERT dan SMOTE untuk Menangani Ketidakseimbangan Kelas. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 256–265. https://doi.org/10.29408/edumatic.v9i1.29666

An, T., Yan, P., Zuo, J., Jin, X., Liu, M., & Wang, J. (2024). Enhancing Cross-Lingual Sarcasm Detection by a Prompt Learning Framework with Data Augmentation and Contrastive Learning. *Electronics (Switzerland)*, *13*(11). https://doi.org/10.3390/electronics13112163

Bayer, M., Kaufhold, M. A., & Reuter, C. (2023). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, *55*(7). https://doi.org/10.1145/3544558

Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2023). A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology*, *32*(4). https://doi.org/10.1145/3583561

Dogra, V., Verma, S., Kavita, Wozniak, M., Shafi, J., & Ijaz, M. F. (2024). Shortcut Learning Explanations for Deep Natural Language Processing: A Survey on Dataset Biases. *IEEE Access*, *12*, 26183–26195. https://doi.org/10.1109/ACCESS.2024.3360306

Fei, H., Chua, T. S., Li, C., Ji, D., Zhang, M., & Ren, Y. (2022). On the Robustness of Aspect-based Sentiment Analysis: Rethinking Model, Data, and Training. *ACM Transactions on Information Systems*, *41*(2). https://doi.org/10.1145/3564281

Gedela, R. T., Baruah, U., & Soni, B. (2024). Deep Contextualised Text Representation and Learning for Sarcasm Detection. *Arabian Journal for Science and Engineering*, *49*(3), 3719–3734. https://doi.org/10.1007/s13369-023-08170-4

Helal, N. A., Hassan, A., Badr, N. L., & Afify, Y. M. (2024). A contextual-based approach for sarcasm detection. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-65217-8

Herawan, D. F., & Saputri, T. R. D. (2025). Benchmarking Model Transformer Modern untuk Analisis Sentimen dan Tren Konsumen dalam Industri Fashion. *Edumatic: Jurnal Pendidikan Informatika*, *9*(3), 945–954. https://doi.org/10.29408/edumatic.v9i3.32657

Hu, Y. H., Liu, T. H., Tsai, C. F., & Lin, Y. J. (2025). Handling Class Imbalanced Data in Sarcasm Detection with Ensemble Oversampling Techniques. *Applied Artificial Intelligence*, *39*(1). https://doi.org/10.1080/08839514.2025.2468534

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R., & Jin, Z. (2023). A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, *5*(10), 1161–1174. https://doi.org/10.1038/s42256-023-00729-y

Javid, B., & Mashayekhi, H. (2025). Classification of imbalanced user reviews using a generative approach. *Social Network Analysis and Mining*, *15*(1). https://doi.org/10.1007/s13278-025-01477-0

Liebeskind, C., & Bączkowska, A. (2025). Sarcastic comments on Reddit and Twitter. *Topics in Linguistics*, *26*(1), 174–193. https://doi.org/10.17846/topling-2025-0008

Liu, H., Yang, B., & Yu, Z. (2024). A Multi-View Interactive Approach for Multimodal Sarcasm Detection in Social Internet of Things with Knowledge Enhancement. *Applied Sciences (Switzerland)*, *14*(5). https://doi.org/10.3390/app14052146

Nasution, A. H., Onan, A., Murakami, Y., Monika, W., & Hanafiah, A. (2025). Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets. *IEEE Access*, *13*, 94009–94025.

https://doi.org/10.1109/ACCESS.2025.3574629

Pandey, R., & Singh, J. P. (2023). BERT-LSTM model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems*, *60*(1), 235–254. https://doi.org/10.1007/s10844-022-00755-z

Šandor, D., & Bagić Babac, M. (2024). Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery*, *52*(2), 213–226. https://doi.org/10.1108/IDD-01-2023-0002

Suhartono, D., Wongso, W., & Tri Handoyo, A. (2024). IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection. *IEEE Access*, *12*, 87323–87332. https://doi.org/10.1109/ACCESS.2024.3416955

Sujana, Y., & Kao, H. Y. (2023). LiDA: Language-Independent Data Augmentation for Text Classification. *IEEE Access*, *11*, 10894–10901. https://doi.org/10.1109/ACCESS.2023.3234019

Sukhavasi, V., & Dondeti, V. (2023). Effective Automated Transformer Model based Sarcasm Detection Using Multilingual Data. *Multimedia Tools and Applications*, *83*(16), 47531–47562. https://doi.org/10.1007/s11042-023-17302-9

Thakkar, G., Preradović, N. M., & Tadić, M. (2024). Examining Sentiment Analysis for Low-Resource Languages with Data Augmentation Techniques. *Eng*, *5*(4), 2920–2942. https://doi.org/10.3390/eng5040152

Zhao, H., Chen, H., Ruggles, T. A., Feng, Y., Singh, D., & Yoon, H. J. (2024). Improving Text Classification with Large Language Model-Based Data Augmentation. *Electronics (Switzerland)*, *13*(13). https://doi.org/10.3390/electronics13132535