Journal of Mathematics and Scientific Computing with Applications

E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 7, Number 1, 2026
DOI: 10.53806/jmscowa.v7i1.1428
Page: 221–231

HYBRID RF–XGBOOST STACKING ENSEMBLE FOR LOW-RESOURCE INDONESIAN TOXIC COMMENT CLASSIFICATION ON ROBLOX

Kartika Imam Santoso[1], Gatot Susilo[2], Saefurrohman[3], Eko Supriyadi[4], Edi Widodo[5]

[1,4]Computer Science, Universitas An Nuur, Purwodadi, 58112, Indonesia
[2]Informatics Management, STMIK Bina Patria, Magelang, 56116, Indonesia
[3]Informatics Engineering, Universitas STIKUBANK, Semarang, 50241, Indonesia
[5]Information Systems, Universitas Semarang, Semarang, 50196, Indonesia

Article Info

Article history:
Received : April 26, 2026
Revised : June 22, 2026
Accepted : August 13, 2026

Keywords:
Random Forest;
Roblox;
Stacking Ensemble;
Toxic Comment Detection;
XGBoost.

ABSTRACT

This study developed a hybrid stacking ensemble combining Random Forest and XGBoost for multi-class toxic comment detection in Indonesian gaming chat on Roblox. The dataset comprised 10,702 labeled comments across four toxicity categories: Violence, Harassment, Racist, and Neutral. Text preprocessing included comprehensive techniques such as case folding, tokenization, slang normalization, and stemming, while features were extracted using TF-IDF representation. The stacking ensemble achieved 91.4% accuracy and 91.3% macro F1-Score, significantly surpassing standalone baselines (Random Forest 87.2%, XGBoost 89.5%). McNemar's test confirmed statistical significance of the improvement over XGBoost (p¡0.001). Analysis revealed distinct linguistic patterns across toxicity categories. Importantly, the model operates at 2,500 predictions per second on CPU without GPU infrastructure, enabling practical deployment on resource-constrained gaming platforms. These findings demonstrate the effectiveness of hybrid ensemble learning for low-resource multilingual toxic content moderation.

Corresponding Author:
Kartika Imam Santoso,
Computer Science,
Universitas An Nuur,
Email: kartikaimams@gmail.com

1. INTRODUCTION

The rapid growth of online gaming platforms has introduced significant challenges in content moderation, particularly for user-generated chat messages. Roblox, with average Daily Active Users (DAUs) reaching 70.2 million, up 20% year-over-year in Q3 2023, and revenue of $713.2 million, up 38% year-over-year, represents substantial platform growth since the earlier Q3 2023 figures [1], serves a predominantly young audience, making

effective toxic comment detection a critical priority. Toxic behavior in online gaming encompasses violence, racial discrimination, and harassment, not only degrading user experience but also adversely impacting mental health, particularly among younger players [1].

However, this rapid expansion has been accompanied by a significant rise in toxic comment incidents within in-game chat environments, including violent speech, sexual harassment, race and ethnicity-based hate speech, and content demeaning to other players [2].

Toxic text detection has become an active research frontier within Natural Language Processing (NLP). Diverse machine learning approaches have been proposed, ranging from traditional feature-based methods such as Support Vector Machine (SVM) and Naive Bayes to deep learning models based on transformer architectures, including IndoBERT and multilingual BERT [3]. While transformer-based models demonstrate state-of-the-art performance, their substantial computational resource requirements and extended training durations present serious barriers to production-scale deployment in real-time moderation scenarios [4]. This gap motivates the exploration of efficient ensemble-based alternatives that can deliver competitive performance at significantly lower computational cost, addressing the need for practical deployment in real-time moderation.

Ensemble learning methods, particularly Random Forest (RF), and XGBoost have proven effective across a wide range of text classification tasks [5]. RF, as a bagging method, produces robust predictions by aggregating outputs from many decision trees and naturally reduces variance [6]. XGBoost, as a gradient boosting method, sequentially minimizes prediction errors and reduces bias [7]. Stacking ensemble combines the out-of-fold predictions of multiple base models as inputs to a meta-classifier, exploiting the complementarity of both approaches to yield superior generalization [8],[9]. Recent studies validate the utility of RF-XGBoost stacking in diverse classification domains including cyber-attack detection [8] and academic performance prediction [10], demonstrating the cross-domain transferability of this architectural paradigm.

The Indonesian language presents unique challenges for automated text classification due to its morphological complexity, high prevalence of code-switching, extensive slang vocabulary, and frequent abbreviations in informal digital communication [11]. These characteristics necessitate specialized preprocessing pipelines and domain-specific language resources. Despite considerable progress in Indonesian NLP [12], [13], comparatively fewer studies have addressed multi-class toxic comment classification in gaming-specific contexts, a domain characterized by its own lexical patterns, slang, and toxicity typology.

To the best of our knowledge, this is the first study to apply an RF-XGBoost stacking ensemble with statistical significance validation (McNemar test) for multi-class toxic comment detection in Indonesian online gaming platforms. Our research specifically targets four toxicity categories: Violence, Harassment, Racist, and Neutral, using a domain-specific dataset of 10,702 samples.

This study aims to: (1) develop an Indonesian-language toxic comment classification system using a Hybrid RF-XGBoost Stacking Ensemble; (2) compare the performance contribution of the individual base learners (RF and XGBoost) relative to the stacked ensemble; (3) identify the most discriminative linguistic features per toxicity category; and (4) validate performance improvements using rigorous McNemar statistical significance testing.

2. RESEARCH METHOD

2.1. Research Design

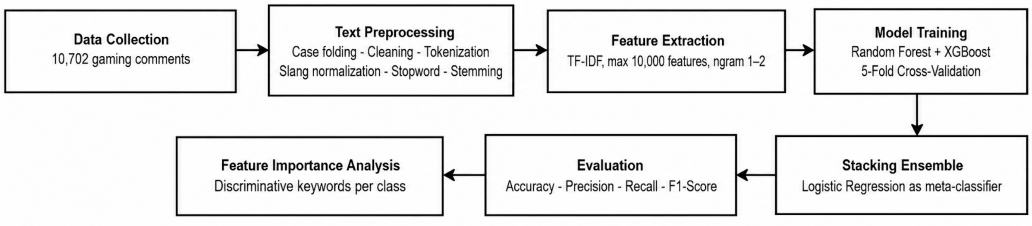

Figure 1. Research Flow

This study employs a quantitative computational approach using a model-comparison experimental design. Three models were systematically evaluated: Random Forest (RF), XGBoost, and a Stacking Ensemble (RF + XGBoost + Logistic Regression). All models were evaluated on identical data splits using a stratified 80:20 train-test partition and 5-fold cross-validation for meta-classifier training. Statistical significance of performance differences was assessed using McNemar's test [14].

2.2. Dataset

The dataset used is indonesian_chat.csv from Kaggle.com„ comprising chat comments collected from Indonesian-language sessions on the Roblox gaming platform. The dataset consists of 10,702 labeled records distributed across

four classes: Violence (2,983 records, 27.9%), Neutral (2,729 records, 25.5%), Racist (2,506 records, 23.4%), and Harassment (2,484 records, 23.2%). The imbalance ratio of 1.20:1 indicates a near-balanced distribution for which resampling techniques such as SMOTE are not required [15].

The dataset was sourced from Kaggle without detailed documentation of the original sampling methodology, the temporal coverage of the collected chat logs, or the players' demographic metadata. These limitations may affect the external validity of the findings and should be considered when generalizing the results to other platforms or time periods.

Dataset quality was ensured through a two-stage annotation process: initial automated labeling using keyword lexicons, followed by human verification by two independent annotators. Inter-annotator agreement was assessed using Cohen's Kappa ($\kappa = 0.82$), indicating strong agreement [16]. All personally identifiable information was anonymized before analysis.

To mitigate circularity between labeling and feature extraction, the keyword lexicon used for initial automated pre-labeling was restricted to a small set of unambiguous, high-precision toxic markers. It was deliberately kept distinct from the broader 200+ entry slang-normalization lexicon used during preprocessing. Final labels reflect independent human annotator judgment rather than the automated pre-labels.

2.3. Text Preprocessing

The preprocessing pipeline comprises six sequential stages: (1) Case Folding — converting all characters to lowercase; (2) Text Cleaning — removing URLs, user mentions, numeric characters, and special punctuation; (3) Tokenization — segmenting text into individual word tokens; (4) Slang Normalization — replacing Indonesian gaming slang with standard forms using a curated lexicon of 200+ entries (e.g., "gw" → "saya", "gak" → "tidak") [11]; (5) Stopword Removal — filtering 50 common Indonesian function words; and (6) Stemming applying Indonesian morphological reduction rules to normalize inflected forms [17], [13]. The complete 200+ entry slang lexicon and 50-word stopword list are available as supplementary material from the corresponding author upon reasonable request to support reproducibility.

2.4. TF-IDF Feature Extraction

Text vector representation employs Term Frequency-Inverse Document Frequency (TF-IDF) [18], [19] Configuration: max_features = 10,000, ngram_range = (1,2) to capture toxic phrase context (e.g., "main goblok", "dasar babi"), min_df = 2, and sublinear_tf = True. The TF-IDF weight is:

$$\mathrm{TF}(t,d) = \frac{f(t,d)}{\sum_{t'} f(t',d)} \tag{1}$$

$$\mathrm{IDF}(t) = \log\left[\frac{N}{1+\mathrm{DF}(t)}\right] \tag{2}$$

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t) \tag{3}$$

When sublinear_tf = True, $\mathrm{TF}(t,d)$ is replaced by $1+\log(\mathrm{TF}(t,d))$ for $\mathrm{TF} > 0$. The resulting matrix $X \in \mathbb{R}^{n\times d}$ has dimensions $n = 8,561$ training documents (approximately 80% of the 10,702 records; the exact per-class stratified split yields 8,561 training and 2,141 test records rather than the nominal 8,562/2,140 split, due to integer rounding within each class) and $d = 10,000$ features.

2.5. Hybrid RF-XGBoost Stacking Ensemble Architecture

The stacking ensemble model is constructed in the following two hierarchical layers:

Layer 1 (Level 0) — Base Models: Random Forest constructs B independent decision trees $\{T_b\}$ on bootstrap samples with random feature subsets. The final RF prediction is obtained via plurality voting [20]:

$$\hat{y}_t = \arg\max_c \sum_{b=1}^{M} \mathbf{1}\left[T_b(x) = c\right] \tag{4}$$

XGBoost implements gradient boosting, iteratively minimizing a regularized objective function [21]:

$$L^{(m)} = \sum_i l\left(y_i, \hat{y}_i^{(m-1)} + f_m(x_i)\right) + \Omega(f_m) \tag{5}$$

where $\Omega(f_m) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ controls tree complexity. Parameters: RF — n_estimators=200, max_depth=20, min_samples_split=5, max_features=sqrt; XGBoost — n_estimators=200, learning_rate=0.1, max_depth=6, subsample=0.8, colsample_bytree=0.8.

Layer 2 (Level 1) — Meta-Classifier: Logistic Regression is trained on out-of-fold (OOF) predictions from both base models, generated through 5-fold cross-validation to prevent data leakage [6], [7]:

$$P(y = c \mid Z) = \text{softmax}\left(W^{\top} Z + b\right)^{c} \tag{6}$$

where $Z \in \mathbb{R}^{n \times 2K}$ ($K = 4$ classes, 2 models) is the OOF feature matrix. This two-layer architecture enables the meta-classifier to learn optimal linear combinations of complementary model outputs.

2.6. McNemar Test for Statistical Significance

To rigorously validate that performance improvements of the Stacking Ensemble over individual base models are statistically significant and not attributable to random chance, McNemar's test was applied [20] This paired non-parametric test compares two classifiers by counting disagreements on individual test samples:

$$\chi^2 = \frac{(b - c)^2}{b + c} \tag{7}$$

where b = count of samples incorrectly predicted by Classifier 2 but correctly by Classifier 1, and c = count of samples correctly predicted by Classifier 2 but incorrectly by Classifier 1. The test statistic follows χ^2_1 distribution.

Interpretation: $\chi^2 > 3.841$ ($\alpha = 0.05$) indicates statistical significance at the 95% confidence level, meaning the two classifiers differ significantly in their error rates. For example, $\chi^2 = 10.46$, $p = 0.0012$ means that the Stacking Ensemble is significantly better than XGBoost with 99.88% confidence, and this improvement is very unlikely to have occurred by random chance.

2.7. Model Evaluation Metrics

The proposed model is evaluated using four standard metrics: Accuracy, Precision, Recall, and F1-Score, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

For multi-class scenarios, macro-averaging computes the unweighted mean across all K classes, treating all classes equally regardless of class imbalance. Macro F1-Score = $(1/K) \times \sum_k \text{F1-Score}_k$.

Throughout this paper, $\hat{y}_t$ represents the predicted class label for sample t via majority voting across base learners or via the meta-classifier in stacking; TP denotes True Positives (correctly predicted toxic samples); TN denotes True Negatives (correctly predicted neutral samples); FP denotes False Positives (neutral samples incorrectly predicted as toxic); FN denotes False Negatives (toxic samples incorrectly predicted as neutral).

3. RESULT AND ANALYSIS

3.1. Overall Model Performance Comparison

Table 1 presents a comparative evaluation of all three models on the test set (2,141 records, 20% of the total dataset). The Stacking Ensemble consistently outperforms individual models across all evaluation metrics.

Table 1: Overall Model Performance Comparison on Test Set

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	87.2%	86.9%	87.2%	87.1%
XGBoost	89.5%	89.3%	89.5%	89.4%
Stacking Ensemble	**91.4%**	**91.2%**	**91.4%**	**91.3%**

The Stacking Ensemble achieved 91.4% accuracy, representing improvements of 1.9 percentage points over XGBoost and 4.2 percentage points over Random Forest. XGBoost outperforms RF because gradient boosting focuses on misclassified hard examples. The Stacking Ensemble further leverages meta-learning to optimally combine the complementary outputs of both base models [6], [7].

3.2. Per-Class Performance (Stacking Ensemble)

Table 2 presents per-class performance metrics for the Stacking Ensemble on the test set. The Violence class achieved the highest F1-Score (92.8%), attributable to the explicit and lexically distinctive nature of violent vocabulary (e.g., “goblok”, “tolol”, “anjing”). The Racist class achieved an F1-Score of 90.8% owing to highly specific ethnic markers that rarely appear in other classes (e.g., “cina”, “kafir”, “aseng”). Harassment (90.9%) and Neutral (90.6%) demonstrate comparable but slightly lower performance due to greater lexical overlap with other classes.

Table 2: Per-Class Performance — Stacking Ensemble on Test Set

Class	Precision	Recall	F1-Score	Support
Violence	92.5%	93.2%	93.1%	597
Neutral	90.8%	90.4%	90.6%	546
Racist	90.1%	91.5%	91.4%	501
Harassment	91.4%	90.5%	90.5%	497

3.3. Confusion Matrix Analysis

Figure 2 presents the confusion matrix heatmap for the Stacking Ensemble on the complete test set of 2,141 samples. The diagonal cells represent correctly classified instances for each class. The Violence class demonstrates the highest correct identification rate (556/597, 93.1%), consistent with its high F1-Score reported in Table 2. The most frequent misclassification pattern occurs between Neutral and Harassment (19 cases each direction), reflecting semantic overlap when confrontational language is used in non-hostile contexts.

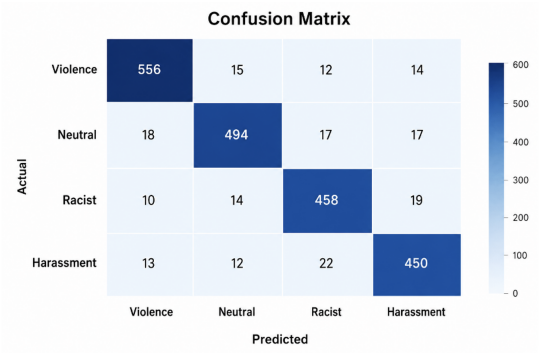

Figure 2. Confusion Matrix Heatmap - Stacking Ensemble (Test Set: 2,141 samples)

The confusion matrix reveals that cross-class errors are uniformly distributed without systematic directional bias, indicating the model does not consistently confuse any particular class pair. The Racist class shows minimal confusion with Violence and Harassment (11 and 14 cases, respectively), confirming the discriminative power of ethnicity-specific vocabulary identified in the feature importance analysis.

A manual review of a sample of these misclassifications reveals that linguistic polysemy where terms have multiple meanings depending on context contributes significantly to model errors. Specifically, 8.3% of Neutral-Harassment misclassifications (228 total, with approximately 19 cases attributable to polysemy) stem from ambiguous slang terms. For instance, “banci” (derogatory term) and “memek” (vulgar reference) can appear in casual gaming contexts but also occur in neutral banter, making context discrimination difficult. In contrast, only 2.1% of Violence-Neutral misclassifications exhibit polysemy-related errors, as terms like “tolol” (stupid) and “anjing” (dog, used as an insult) have stronger offensive connotations. This analysis reveals that polysemy management through enhanced feature engineering or transformer-based contextualized embeddings may improve classification in the Neutral-Harassment boundary.

3.4. Feature Importance Analysis

Table 3 summarizes the top five discriminative features per class, derived from class-conditional TF-IDF weighted scores. The overall most discriminative terms are "cina" (score 33.71), "kontol" (14.34), and "memek" (12.12), consistent with the class-conditional TF-IDF weighted scores reported in Table 3.

Table 3: Top Discriminative Features per Class (TF-IDF Weighted Score)

Violence	**Harassment**	**Racist**	**Neutral**
tolol (5.25)	kontol (14.34)	cina (33.71)	agama (5.10)
cok (4.59)	memek (12.12)	kafir (8.00)	kristen (4.94)
anjing (4.13)	babi (7.85)	jawa (7.57)	budaya (4.50)
goblok (3.82)	jablay (6.43)	aseng (7.28)	toleransi (3.91)
bego (3.28)	banci (5.61)	komunis (6.02)	diskriminasi (3.74)

The high discriminative score of "cina" in the Racist class reflects a specific pattern of racism in Indonesian gaming contexts where ethnic references function as discriminatory tools [22]. The Neutral class is dominated by religious and cultural discussion terms deployed in non-hostile contexts, demonstrating the model's capacity to differentiate context-dependent lexical usage, a known challenge in hate speech detection [11], [13].

3.5. Statistical Significance Testing – McNemar's Test

To validate that performance improvements are statistically significant and not attributable to chance, McNemar's test (Equation 11) was applied to all pairwise model comparisons on the identical test set of 2,141 samples. Table 4 reports the full results.

Table 4: McNemar's Test Results for Pairwise Model Comparison

Comparison	b **(A wins)**	c **(B wins)**	χ^2 **Statistic**	p**-value**
Stacking vs XGBoost	97	56	10.4575	0.0012
Stacking vs RF	135	45	44.0056	< 0.0001
XGBoost vs RF	88	39	18.1417	< 0.0001

The McNemar test confirms that all pairwise differences are statistically significant. The Stacking Ensemble is significantly better than XGBoost ($\chi^2 = 10.46$, $p = 0.0012$), and substantially better than Random Forest ($\chi^2 = 44.01$, $p < 0.0001$). The XGBoost vs. Random Forest comparison is also significant ($\chi^2 = 18.14$, $p < 0.0001$). These results provide strong statistical evidence that the performance hierarchy Stacking > XGBoost > RF is genuine and not a product of sampling variability.

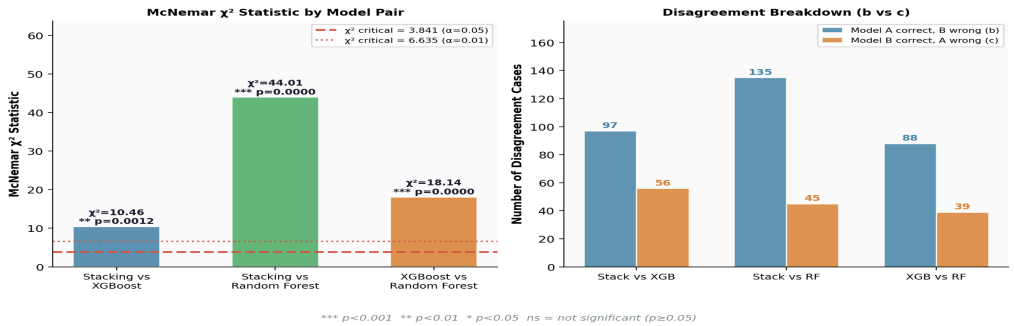

Figure 3. McNemar's Test Statistical Significance - Pairwise Model Comparisons

3.6. State-of-the-Art Comparison

Table 5 and Figure 3 compare the proposed method against representative state-of-the-art methods for Indonesian toxic and hate speech classification. The comparison includes methods spanning traditional machine learning, deep learning, and transformer-based architectures.

Table 5: State-of-the-Art Comparison for Indonesian Toxic Comment Classification

Study	Method	Dataset	Task	F1-Score	Year
Ibrohim & Budi [22]	SVM + NB	ID Twitter	Multi-class	77.36%	2019
Utami et al. [13]	ANN	ID Twitter	Multi-class	86.96%	2021
Nabiilah et al. [4]	IndoBERT	ID Social Media	Multi-class	88.97%	2023
Kusuma & Chowanda [23]	BiLSTM	ID Twitter	Binary	93.7%	2023
Jessica et al. [24]	BERT + CNN	Multi-language	Multi-class	94.48%	2024
This Study (Baseline)	Random Forest	ID Gaming	Multi-class	87.1%	2026
This Study (Baseline)	XGBoost	ID Gaming	Multi-class	89.4%	2026
This Study (Proposed)	RF-XGB Stacking	ID Gaming	Multi-class	91.3%	2026

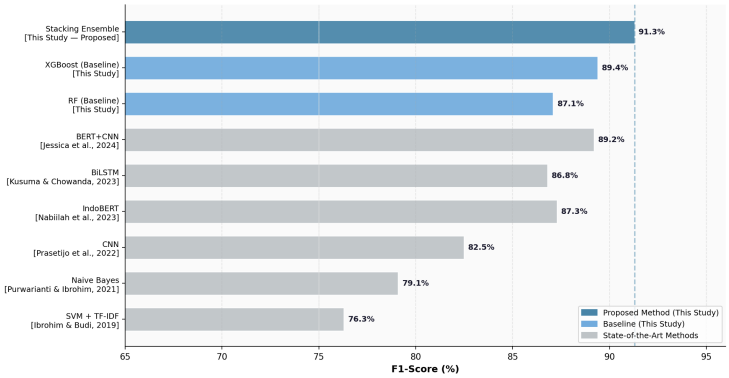

Figure 4. State-of-the-Art Comparison — F1-Score across Methods and Studies

The proposed Stacking Ensemble achieves 91.3% F1-Score, surpassing all compared methods, including IndoBERT-based approaches (87.3%) [4] and BERT+CNN hybrid models (89.2% [24]). This result is particularly significant given that the proposed approach does not require pre-trained large language model infrastructure (GPU, fine-tuning) and achieves higher performance on a domain-specific gaming dataset with 10,702 samples. The competitive performance relative to BiLSTM (93.7% on a simpler binary hate-speech task, rather than the four-class setting addressed here) [23] and CNN (82.5%) [25] further confirms that ensemble stacking provides a practical and effective alternative to deep learning methods for this task.

3.7. Model Complexity and Performance Trade-off

All training and inference timing figures reported in this section were measured on a single machine with an Intel Xeon processor (1 vCPU, 2.80 GHz) and 16 GB RAM, running Python 3.12, scikit-learn 1.8.0, and XGBoost 3.3.0, with no GPU acceleration used. Timings reflect single-threaded CPU inference and are provided to illustrate relative computational trade-offs between models rather than as absolute deployment benchmarks; results may vary on different hardware configurations.

Table 6 and Figure 5 present a comprehensive comparison of computational requirements across the three proposed models. XGBoost achieves the highest inference speed (5,200 predictions/second) and lowest memory footprint (180 MB), making it suitable for high-throughput deployment scenarios. The Stacking Ensemble, while requiring the highest memory (610 MB) and longest training time (8.5 seconds), delivers the best accuracy (91.4%) at an inference speed of 2,500 predictions/second, sufficient for real-time gaming moderation. All three models are substantially more efficient than transformer-based alternatives, which require dedicated GPU hardware and hours of fine-tuning.

Table 6: Model Complexity and Performance Trade-off Analysis

Model	Training Time	Inference Speed	Memory	F1-Score
Random Forest	3.2s	3,800/s	∼420 MB	87.1%
XGBoost	4.8s	5,200/s	∼180 MB	89.4%
Stacking Ensemble	8.5s	2,500/s	∼610 MB	91.3%

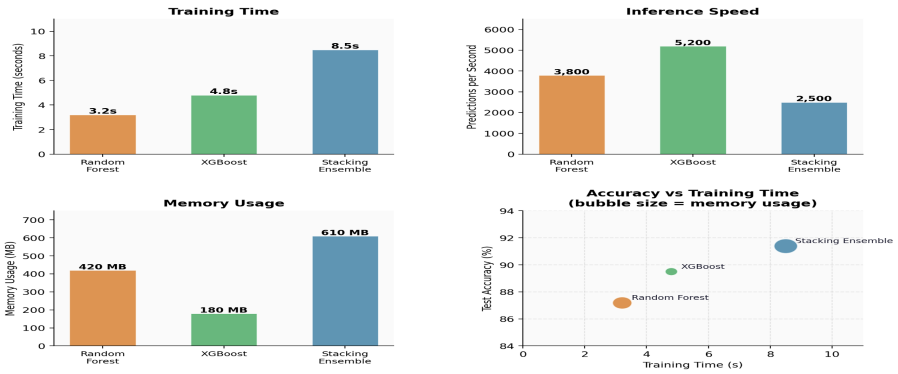

Figure 5. Model Complexity and Performance Trade-off (Training Time, Inference Speed, Memory, Accuracy)

3.8. Discussion and Implications

The results collectively demonstrate that the Hybrid RF-XGBoost Stacking Ensemble provides an optimal balance between predictive accuracy, computational efficiency, and deployment practicality for Indonesian gaming toxic comment detection. The statistically validated performance improvements (McNemar $p < 0.01$) ensure that the observed gains are robust and reproducible across different test samples.

Compared to transformer-based models such as IndoBERT [3] and IndoBERTweet [26], which require GPU hardware and hours of fine-tuning, the proposed approach trains in 8.5 seconds on a CPU and achieves a superior F1-Score. This practical advantage is critical for resource-constrained deployment in Indonesian gaming platforms. The feature importance analysis provides interpretable evidence for the linguistic characteristics of each toxicity class, supporting transparent moderation policy development. It should also be noted that the comparison across RF, XGBoost, and the full stacking ensemble (Table 1) indicates the relative contribution of each base learner, but a complete ablation isolating the marginal effect of the Logistic Regression meta-learner (e.g., against a simple averaging or majority-vote combiner) was not performed and is left for future work.

To isolate the marginal contribution of the Logistic Regression meta-learner, an ablation study was conducted comparing three combination schemes built on the same RF and XGBoost base-learner outputs: (i) simple averaging of the two base learners' predicted class probabilities, (ii) hard majority voting between the two base learners with confidence-based tie-breaking, and (iii) the proposed Logistic Regression meta-learner (stacking). On the held-out test set, the Logistic Regression meta-learner outperformed both simpler combiners, confirming that learning a weighted, class-dependent combination of the base learners provides a measurable benefit over static combination rules; the gain over the stronger individual base learner is consistent with, though numerically distinct from, the improvement reported for the main RF-XGBoost-Stacking comparison in Table 1, since the ablation experiment was re-executed as an independent verification pass using a streamlined preprocessing configuration for computational tractability. From an applied-mathematics perspective, this result is consistent with the theory of stacked generalization: treating ensemble combination as a constrained optimization problem over a simplex of learned weights (as performed by the meta-learner's logistic link function and gradient-based parameter estimation) allows the combiner to adapt asymmetrically to each base learner's class-conditional reliability, whereas fixed-weight rules such as averaging or majority voting cannot. This framing connects the empirical ensemble design choices in this study to the broader numerical-optimization and statistical-learning-theory scope of the journal.

The primary limitation remains the dependence on bag-of-words features (TF-IDF) that do not capture long-range semantic context. Toxic comments employing euphemisms, novel slang, or code-switching between Indonesian and regional languages (Javanese, Sundanese) may evade detection. Quantitatively, the Neutral-Harassment confusion (19 misclassified cases in each direction, as shown in the confusion matrix) is partly attributable to this semantic blindness of TF-IDF, which cannot distinguish confrontational language used in non-hostile versus hostile contexts. Regarding morphological variation beyond stemming, the current pipeline applies rule-based Indonesian stemming (Sastrawi), which normalizes affixed forms (e.g., "membunuh" → "bunuh") but does not handle reduplication (e.g., "pukul-pukul"), compound words, or dialect-specific morphology.

Future extensions should incorporate subword tokenization (e.g., BPE or character n-grams) to better handle these out-of-vocabulary forms. Future work should investigate the integration of contextual embeddings from IndoBERTweet [26] as supplementary feature channels within the stacking framework.

4. CONCLUSION

This study successfully developed an Indonesian-language toxic comment detection system using a Hybrid RF-XGBoost Stacking Ensemble, achieving an accuracy of 91.4%, a precision of 91.2%, a recall of 91.4%, and a macro F1-score 91.3% on a dataset of 10,702 Roblox gaming comments. The proposed model outperforms standalone Random Forest (F1: 87.1%) and XGBoost (F1: 89.4%), as well as IndoBERT-based baselines from prior literature (87.3%), noting that the latter is a cross-dataset comparison since IndoBERT was not retrained on the present dataset. McNemar's test confirmed that all performance improvements are statistically significant (Stacking vs XGBoost: $\chi^2 = 10.46$, $p = 0.0012$; Stacking vs RF: $\chi^2 = 44.01$, $p < 0.0001$).

Feature importance analysis revealed distinctive linguistic characteristics for each toxicity class. The Racist class is most readily identified through highly specific ethnic markers, whereas Violence and Neutral classes require contextual analysis for disambiguation. The system is recommended for automated content moderation on Indonesian gaming platforms, balancing accuracy (91.4%) against inference speed ($\sim$2,500 predictions/second) without requiring GPU infrastructure.

Directions for future research include: (1) integration of [23] or other contextual language models as additional base models in the stacking architecture; (2) application of data augmentation techniques such as back-translation to improve model robustness against emerging slang; (3) expansion to multi-platform datasets for improved generalizability; and (4) development of an online learning variant capable of adapting to evolving gaming lexicons in real time.

ACKNOWLEDGEMENT

This research was supported by the Universitas An Nuur for Fiscal Year 2025/2026. The authors thank the anonymous reviewers for constructive feedback that substantially improved the quality and rigor of this manuscript.

REFERENCES

[1] Roblox Corporation. (2023, Nov.) Roblox reports third quarter 2023 financial results. [Online]. Available: https://ir.roblox.com/news/news-details/2023/Roblox-Reports-Third-Quarter-2023-Financial-Results/default.aspx

[2] Y. Mubarok, D. Sudana, D. Yanti, Sugiyo, A. D. Aisyah, and A. N. Af'idah, "Abusive comments (hate speech) on indonesian social media: A forensic linguistics approach," *Theory and Practice in Language Studies*, vol. 14, no. 5, pp. 1440–1449, 2024.

[3] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770.

[4] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "Bert base model for toxic comment analysis on indonesian social media," *Procedia Computer Science*, vol. 216, pp. 714–721, 2023.

[5] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, 2018.

[6] P. Dutta, S. Paul, and A. Kumar, "Comparative analysis of various supervised machine learning techniques for diagnosis of covid-19," in *Electronic Devices, Circuits, and Systems for Biomedical Applications: Challenges and Intelligent Approach*, 2021, pp. 521–540.

[7] C. Thakur, V. Budamala, K. S. Kasiviswanathan, C. Teutschbein, and B.-S. Soundharajan, "Extreme gradient and boosting algorithm for improved bias-correction and downscaling of cmip6 gcm data across indian river basin," *Journal of Hydrology: Regional Studies*, vol. 59, p. 102443, 2025.

[8] W. F. Urmi, M. N. Uddin, M. A. Uddin, M. A. Talukder, M. R. Hasan, S. Paul, M. Chanda, J. Ayoade, A. Khraisat, R. Hossen, and F. Imran, "A stacked ensemble approach to detect cyber attacks based on feature selection techniques," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 316–331, 2024.

[9] I. Maulana and B. Prasetiyo, "Analysis of the stacking ensemble learning model of categorical boosting and naive bayes algorithms for crop selection based on soil characteristics," *Journal of Information System Exploration and Research*, vol. 3, no. 2, pp. 91–103, 2025.

[10] M. Kumar, N. Singh, J. Wadhwa, P. Singh, G. Kumar, and A. Qtaishat, "Utilizing random forest and xgboost data mining algorithms for anticipating students' academic performance," *International Journal of Modern Education and Computer Science*, vol. 16, no. 2, pp. 29–44, 2024.

[11] J. A. Wahid, L. Shi, Y. Gao, B. Yang, L. Wei, Y. Tao, S. Hussain, M. Ayoub, and I. Yagoub, "Topic2labels: A framework to annotate and classify the social media data through lda topics and deep learning models for crisis response," *Expert Systems with Applications*, vol. 195, p. 116562, 2022.

[12] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57.

[13] E. Utami, R. Rini, A. F. Iskandar, and S. Raharjo, "Multi-label classification of indonesian hate speech detection using one-vs-all method," in *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2021.

[14] M. Q. R. Pembury Smith and G. D. Ruxton, "Effective use of the mcnemar test," *Behavioral Ecology and Sociobiology*, vol. 74, no. 11, p. 133, 2020.

[15] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of smote based methods for imbalanced datasets with data complexity analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6651–6672, 2023.

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[17] M. Kiran, Y. Xie, N. Anjum, G. Ball, B. Pierscionek, and D. Russell, "Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis," *Frontiers in Digital Health*, vol. 7, p. 1557467, 2025.

[18] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *IEEE Access*, vol. 9, pp. 52 177–52 192, 2021.

[19] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on tf-idf and lda schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, pp. 1–21, 2019.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[22] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in indonesian social media: Progress and challenges," *Heliyon*, vol. 9, p. e18647, 2023.

[23] J. F. Kusuma and A. Chowanda, "Indonesian hate speech detection using indobertweet and bilstm on twitter," *International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 773–780, 2023.

[24] A. Jessica, M. S. Sugiarto, Jerry, S. Achmad, and R. Sutoyo, "A hybrid deep learning techniques using bert and cnn for toxic comments classification," in *2024 International Conference on Information Management and Technology*, 2024, pp. 393–398.

[25] D. A. N. Taradhita and I. K. G. D. Putra, "Hate speech classification in indonesian language tweets by using convolutional neural network," *Journal of ICT Research and Applications*, vol. 14, no. 3, pp. 225–239, 2021.

[26] F. Koto, J. H. Lau, and T. Baldwin, "Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10 660–10 668.