A Lightweight Machine Learning Model for Early Detection of Cyberbullying in Online Gaming Communities to Support Digital Character Education

Farhan Badrani[1]; Nuur Wachid Abdul Majid[2*]

[1,2]*Information Systems and Technology Education, Universitas Pendidikan Indonesia, Indonesia*
[2*]*Corresponding Email: nuurwachid@upi.edu*

Article History:
Received: Dec 22, 2025
Revised: Jan 26, 2026
Accepted: Feb 01, 2026
Online First: Feb 27, 2026

Keywords:
Cyberbullying,
Digital Character education,
Digital Literacy,
Linear SVM,
Slang Normalization.

Kata Kunci:
Linear SVM,
Literasi digital,
Normalisasi slang,
Pendidikan Karakter Digital,
Perundungan Siber.

Abstract: This study develops a lightweight early-warning model to identify toxic utterances as practical indicators of cyberbullying in Indonesian-language conversations within the Roblox gaming community, to support digital character education and child online safety. A corpus of 2,798 publicly available comments was manually annotated into Safe and Toxic categories and divided into training and testing sets. Text preprocessing included case folding, noise removal, tokenization, Roblox-specific slang normalization, stemming, and stopword removal. Text features were represented using term frequency–inverse document frequency (TF-IDF) unigram–bigram vectors. A linear Support Vector Machine (SVM) was evaluated against Multinomial Naïve Bayes as a baseline model. Results from hold-out testing indicate that the SVM achieved 82.14% accuracy and a macro-F1 score of 0.82, outperforming the baseline. Cross-validation results show performance variability, highlighting the need for continuous updates of domain-specific slang resources and broader data coverage. From an educational perspective, the proposed prototype can function as a non-punitive screening tool to support digital literacy instruction, school counselling, and parental mediation within a human-in-the-loop framework.

Abstrak: Penelitian ini mengembangkan model peringatan dini yang ringan untuk mengidentifikasi ujaran toxic sebagai indikator praktis cyberbullying dalam percakapan berbahasa Indonesia pada komunitas game Roblox, dengan tujuan mendukung pendidikan karakter digital dan keamanan anak di ruang daring. Penelitian menggunakan korpus berisi 2.798 komentar publik yang dianotasi secara manual ke dalam kategori Aman dan Toxic, kemudian dibagi ke dalam data latih dan data uji. Tahapan praproses meliputi case folding, pembersihan noise, tokenisasi, normalisasi slang berbasis kamus khusus Roblox, stemming, dan penghapusan stopword. Representasi fitur teks dilakukan menggunakan term frequency–inverse document frequency (TF-IDF) unigram–bigram. Kinerja model Support Vector Machine (SVM) linear dibandingkan dengan Multinomial Naive Bayes sebagai baseline. Hasil pengujian menunjukkan bahwa SVM mencapai akurasi 82,14% dan macro-F1 sebesar 0,82, lebih tinggi dibandingkan model baseline. Validasi silang menunjukkan adanya variasi performa, yang mengindikasikan perlunya pembaruan kamus slang dan perluasan cakupan data. Dari perspektif pendidikan, prototipe yang dikembangkan berpotensi digunakan sebagai alat skrining non-punitif untuk mendukung pembelajaran literasi digital, layanan bimbingan dan konseling, serta mediasi orang tua dengan prinsip human-in-the-loop.

How to cite:
Badrani, F., & Majid, N. W. A. (2026). A Lightweight Machine Learning Model for Early Detection of Cyberbullying in Online Gaming Communities to Support Digital Character Education. *Edunesia : Jurnal Ilmiah Pendidikan, 7*(2), 876-892.

https://doi.org/10.51276/edu.v7i2.1650

A. Introduction

Social interactions among children and adolescents increasingly take place in digital environments, and online gaming communities have become a central arena of contemporary youth culture. Alongside opportunities for collaboration and peer bonding, real-time and competitive communication in games may also trigger mockery, insults, and other forms of cyberbullying that harm psychological well-being and disrupt healthy social climates, with longitudinal evidence confirming that such experiences are associated with lasting negative psychosocial outcomes among youth (Marciano et al., 2020). From an educational standpoint, cyberbullying is not only a safety issue; it is closely connected to digital character education, digital literacy, and schools' efforts to nurture responsible online behavior (Isnawan, 2025). Evidence from prevention research suggests that effective cyberbullying interventions are typically multi-component, combining education, reinforcement of positive norms, and coordinated support across schools, families, and communities. In this context, educators and parents need practical support tools that enable more responsive monitoring without replacing professional judgement, for example, non-punitive early screening mechanisms that function as prompts for guidance and counselling rather than instruments for punishment (Polanin et al., 2022; Tozzo et al., 2022). Reviews of technology-based approaches further emphasize that detection tools can support prevention when designed to complement human decision-making and used transparently within educational ecosystems (Chan et al., 2023).

Online gaming settings introduce additional challenges because toxic interactions often emerge through competitive "trash talk," performance-based provocation, and fast-paced exchanges where meaning depends heavily on context (Hu et al., 2025). In Indonesian Roblox-related communities, language use is highly informal and dynamic: users frequently employ abbreviations, creative spellings to evade filters, and code-mixing between Indonesian and English gaming terms. These linguistic patterns complicate manual supervision by adults and reduce the effectiveness of generic keyword-based filtering.

Prior research on cyberbullying and abusive language detection has produced a wide range of methods and datasets, including broad systematic reviews of machine learning-based cyberbullying detection (Balakrisnan & Kaity, 2023), domain-specific reviews in gaming contexts (Hu et al., 2025), emerging toxicity datasets for gaming communities (Naseem et al., 2025), and machine learning-based verbal harassment detection specifically designed for online game environments (Hibatullah et al., 2025). In Indonesia, most computational studies still concentrate on general social media platforms, including hate speech detection in Indonesian tweets (Elisabeth et al., 2020; Ibrohim & Budi, 2023; Bustamin et al., 2025) and abusive language detection for regional languages such as Javanese and Sundanese (Putri et al., 2021), while domain-focused investigations for gaming discourse remain limited. At the same time, Indonesian NLP studies consistently report that orthographic variation, slang, and code-mixing can degrade classification performance when not properly handled through normalization and domain-aware preprocessing (Ibrohim & Budi, 2023; Bustamin et al., 2025).

https://doi.org/10.51276/edu.v7i2.1650

Despite these advances, two gaps remain important for education-oriented applications. First, there is a lack of publicly described Indonesian-language corpora that specifically capture toxicity in gaming-related discourse and can be used for transparent experimentation and classroom or counselling prototypes. Second, many studies emphasize complex architectures, while evidence remains limited on how far a lightweight, interpretable pipeline augmented with domain-specific slang normalization can provide reliable early-warning signals that are feasible to maintain in resource-constrained educational settings.

To address the gaps above, this study integrates a Roblox-specific slang lexicon (420 entries) into an Indonesian text-processing pipeline. It evaluates a lightweight term frequency–inverse document frequency (TF-IDF) representation with a linear Support Vector Machine classifier, using Multinomial Naïve Bayes as a baseline. Beyond reporting technical performance, the study frames the model as an educational support tool. It provides a simple prototype interface to encourage transparent and responsible use in line with a human-in-the-loop principle.

Accordingly, this study addresses two research questions: (RQ1) What linguistic characteristics and patterns of toxic utterances appear in Indonesian Roblox-themed community conversations?. (RQ2) To what extent can a model that combines Roblox-specific slang normalization, TF-IDF features, and a linear Support Vector Machine effectively distinguish Safe and Toxic comments in this corpus?. Based on these questions, the study aims to (1) characterize toxicity patterns in the dataset, (2) build and evaluate a lightweight detection model, and (3) discuss how the model can be used to support digital character education through non-punitive screening and guided intervention.

B. Method

This investigation employs a quantitative-experimental research design, utilizing supervised machine learning methodologies for text classification, specifically designed to systematically distinguish between Safe comments (labeled 0) and Toxic comments (labeled 1). Figure 1 provides a comprehensive visual representation of the research workflow, encompassing all procedural stages from initial data collection to final model deployment and validation.

https://doi.org/10.51276/edu.v7i2.1650

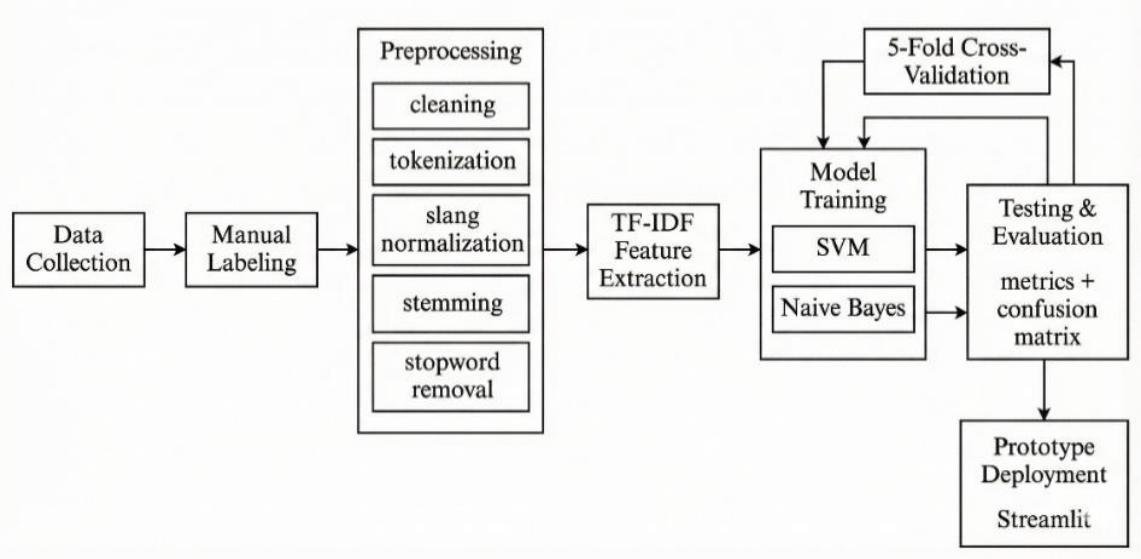

Figure 1. Research Flow of Toxic Speech Detection Pipeline

The fundamental unit of analysis comprises publicly accessible comments on Roblox gaming content, composed exclusively in Indonesian and collected from diverse external community spaces, including comment sections and dedicated community channels on platforms frequented by Indonesian Roblox players. The complete dataset encompasses 2,798 distinct comment samples systematically collected to represent diverse communication patterns, emotional expressions, and interaction styles characteristic of the gaming community. Data partitioning follows a conventional 80-20 split, allocating 2,238 samples for comprehensive model training and reserving 560 for independent testing, and consistently using random state 42 throughout all experimental procedures to ensure complete reproducibility of results and facilitate meaningful comparative assessment across methodological variations.

To rigorously maintain ethical research standards specifically applicable to online data collection contexts, all personal identifiers, including usernames, user IDs, and other potentially identifying metadata, were systematically removed from the dataset through automated anonymization procedures, and the research protocol deliberately avoids verbatim republication of sensitive content categories, thereby comprehensively protecting user privacy while simultaneously enabling legitimate academic investigation of communication patterns. This ethical approach aligns with contemporary standards for research involving human subjects in digital environments and ensures compliance with data protection principles.

Data labeling was meticulously conducted through manual annotation, using a carefully defined binary classification scheme designed to capture essential toxicity dimensions. Comments were systematically assigned to the Toxic category (label 1) when they demonstrably contained insults directed at individuals or groups, harsh or profane language, demeaning utterances designed to belittle others, attacking statements targeting personal characteristics, provocative expressions intended to instigate conflict, or aggressive

https://doi.org/10.51276/edu.v7i2.1650

communications that could constitute recognizable cyberbullying manifestations. Conversely, comments were classified as Safe (label 0) when they exhibited characteristics such as a neutral, informational tone, factual content, expressions of positive support toward community members, or humorous exchanges without identifiable insulting targets or malicious intent.

The annotation process strictly adhered to standardized annotation guidelines explicitly formulated based on established cyberbullying indicators extensively documented in prior scholarly literature, encompassing categories such as personal attacks against individuals, group-based harassment targeting demographic characteristics, aggressive provocations designed to escalate conflicts, and deployment of offensive language intended to cause emotional harm. Although the current study pragmatically employed a single trained annotator for practical feasibility, the annotation protocol was comprehensively documented in written guidelines to enable future implementation of inter-rater reliability assessments and potential expansion to multi-annotator frameworks incorporating inter-coder agreement measurements in subsequent research phases.

The preprocessing pipeline systematically comprises six carefully sequenced stages, specifically designed to transform raw comment text into clean, normalized linguistic representations optimally suited for subsequent feature extraction. Stage one implements comprehensive case folding operations, systematically converting all alphabetic characters to lowercase format, thereby substantially reducing vocabulary size and effectively eliminating case-based lexical variations that do not contribute to semantic meaning. Stage two executes thorough noise-cleaning procedures, systematically removing URLs, emoji symbols, user mention tags, hashtag markers, and special characters that do not contribute substantially to semantic value in toxicity classification tasks. Stage three performs tokenization using the widely adopted Natural Language Toolkit to segment processed text into discrete token units suitable for computational analysis.

Stage four constitutes the methodologically critical slang normalization phase, systematically employing the custom-compiled Roblox-specific dictionary containing 420 carefully curated entries to systematically map informal gaming slang expressions, creative orthographic variations, and community-specific terminology to their corresponding standard Indonesian language equivalents, thereby substantially reducing lexical variation and meaningfully enhancing feature consistency across the corpus. Stage five methodically applies Sastrawi stemming algorithms, specifically optimized for Indonesian morphology, to reduce inflected words to their morphological roots, thereby further consolidating morphological variants and reducing vocabulary complexity. Finally, stage six implements comprehensive stopword removal procedures, using Sastrawi-curated stopword lists, to systematically eliminate high-frequency function words that carry minimal discriminative value for classification while preserving content-bearing lexical items essential for meaningful semantic analysis (Bustamin et al., 2025).

Feature representation systematically employs the Term Frequency-Inverse Document Frequency vectorization methodology, configured with carefully optimized

https://doi.org/10.51276/edu.v7i2.1650

parameters that balance computational efficiency with comprehensive vocabulary coverage. The maximum features parameter was deliberately limited to 5,000 dimensions to maintain computational tractability while ensuring sufficient capacity to capture domain-specific vocabulary richness. The n-gram range was strategically configured to include both unigrams and bigrams, thereby capturing both individual term frequencies and local contextual patterns encoded in adjacent word pairs. A minimum document frequency of 2 was established to systematically filter exceedingly rare terms that are likely noise rather than meaningful signals. The maximum document frequency was set to 0.95 to systematically eliminate overly common terms that appear ubiquitously across documents and thus lack discriminative power for classification. This carefully calibrated configuration yielded 1,267 effective features across the entire corpus, representing a well-balanced vocabulary space that captures domain-specific terminology while maintaining robust generalization capacity.

The primary classification model employs a Linear Support Vector Machine with a regularization parameter C = 1.0 and a linear kernel. Support Vector Machine methodology was deliberately selected for its well-documented effectiveness in managing high-dimensional, sparse text representations and for its demonstrated robust generalization across diverse text classification tasks, as comprehensively established in the foundational machine learning literature (Cortes & Vapnik, 1995; Joachims, 1998). The comparative baseline model uses a Multinomial Naïve Bayes architecture with a Laplace smoothing parameter alpha set to 1.0, a probabilistic classification approach widely deployed in text classification applications, and a standard performance benchmark for comparison.

Model performance assessment systematically employs multiple complementary evaluation metrics, providing a comprehensive characterization of classification quality across diverse aspects. Accuracy quantifies overall classification correctness as the proportion of correct predictions across all test samples. Precision measures the proportion of true positives among all positive predictions generated by the model, thereby indicating the system's ability to avoid false-positive alerts. Recall quantifies the proportion of actual positive cases that were correctly identified by the model, thereby reflecting detection completeness and sensitivity. F1-score computes the harmonic mean of precision and recall, thereby providing a balanced composite metric particularly valuable for datasets with relatively balanced class distributions, such as the present corpus.

Confusion matrix analysis provides a granular breakdown of classification outcomes into true positives, true negatives, false positives, and false negatives, thereby enabling detailed identification of specific error patterns and systematic biases. Beyond conventional hold-out test evaluation conducted on the strategically reserved 20% test set, this investigation implements rigorous 5-fold cross-validation procedures to comprehensively assess model stability and generalization capacity across different data partitions, thereby providing more robust and reliable performance estimates that explicitly account for potential sampling variability and data distribution variations that might influence single-partition evaluation results.

https://doi.org/10.51276/edu.v7i2.1650

Implementation Environment

All experimental procedures were systematically conducted using the Python 3.x programming environment, selected for its extensive machine learning ecosystem and widespread adoption in academic research. Machine learning model development, training, and systematic evaluation used the scikit-learn library, which provides comprehensive, well-documented implementations of Support Vector Machine algorithms, Naïve Bayes classifiers, TF-IDF vectorization, and extensive evaluation metric calculations. The Natural Language Toolkit facilitated efficient tokenization. Sastrawi library, specifically designed and optimized for Indonesian language processing requirements, provided essential support for morphological stemming and stopword removal, tailored to Indonesian linguistic characteristics. The preliminary interactive prototype interface was systematically built using the Streamlit framework, enabling rapid development of a user-friendly demonstration application, particularly suitable for educational deployment contexts and stakeholder engagement activities. This carefully selected technology stack was deliberately chosen to optimally balance robust academic research capabilities with practical deployment feasibility in potentially resource-constrained educational institutional settings.

C. Result

This section systematically presents comprehensive statistical summaries of dataset characteristics, detailed model configuration specifications, and complete documentation of evaluation outcomes. The dataset exhibits a notably balanced class distribution between Safe and Toxic categories, with 50.68% classified as Safe and 49.32% as Toxic, thereby rendering the F1-score a particularly relevant and informative performance indicator alongside standard accuracy metrics for comprehensive performance assessment.

Table 1. Dataset Statistics

Parameter	Value	Description
Total Sample	2,798	Indonesian-language comments themed around Roblox
Training (80%)	2,238	Used for model training
Testing (20%)	560	Used for hold-out evaluation
Safe Class (0)	1.418 (50.68%)	Non-toxic label
Toxic Class (1)	1.380 (49.32%)	Toxic/cyberbullying indicator label
Slang Dictionary	420 entries	Normalization of community terms/slang

Table 1 summarizes the corpus characteristics. The dataset contains 2,798 Indonesian-language Roblox-related comments with a nearly balanced distribution between Safe (50.68%) and Toxic (49.32%) labels. The 420-entry slang lexicon reflects the amount of domain-specific vocabulary handled during preprocessing.

https://doi.org/10.51276/edu.v7i2.1650

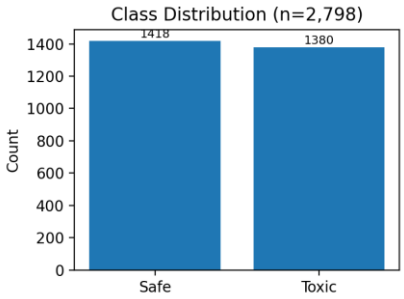

Figure 2. Class Distribution of the Roblox Toxicity Dataset

Figure 2 shows the near balance between Safe and Toxic labels. This balanced distribution helps prevent accuracy from being driven by a dominant class and supports the use of macro-averaged metrics.

Table 2. Preprocessing Pipeline

Stage	Method
1	Case folding (lowercase)
2	Cleaning (URLs, emojis, mentions, hashtags, special characters)
3	Slang normalization (420-entry lexicon)
4	Tokenization (NLTK)
5	Stemming (Sastrawi)
6	Stopword removal (Sastrawi)

As listed in Table 2, six preprocessing stages were applied to standardize informal gaming text before feature extraction, including a dedicated slang-normalization step that maps community terms and creative spellings to more consistent forms.

Table 3. Feature Extraction and Model Configuration

Component	Parameter	Value
TF-IDF	max_features	5000
TF-IDF	n-gram range	(1,2)
TF-IDF	min_df	2
TF-IDF	max_df	0.95
Model	SVM kernel	linear
Model	SVM C	1.0
Baseline	Naïve Bayes alpha	1.0
Experiment	random_state	42
Results	actual number of features	1.267

Table 3 details the experimental configuration used consistently across models. Under these TF-IDF settings, the vectorizer produced 1.267 effective features from the corpus, balancing vocabulary coverage and computational efficiency.

https://doi.org/10.51276/edu.v7i2.1650

Table 4. Model Performance on Hold-out Test Set (n=560)

Model	Accuracy	Precision	Recall	F1-score
Linear SVM	0.8214	0.8216	0.8214	0.8214
Naïve Bayes	0.8018	0.8036	0.8018	0.8013

Table 4 reports the hold-out evaluation. The Linear SVM achieved 0.8214 accuracy and 0.8214 macro-F1, outperforming Multinomial Naïve Bayes by 1.96 percentage points in accuracy and by 0.0201 in macro-F1.

Table 5. Confusion Matrix (Linear SVM)

	Predicted: Safe	Predicted: Toxic
Actual: Safe	231 (TN)	53 (FP)
Actual: Toxic	47 (FN)	229 (TP)

Table 5 shows the distribution of prediction outcomes. Misclassifications were relatively symmetric, with 53 false positives (Safe predicted as Toxic) and 47 false negatives (Toxic predicted as Safe).

Figure 3 provides a visual overview of the confusion matrix and shows that both false-positive and false-negative errors occur at comparable rates in the current setting.

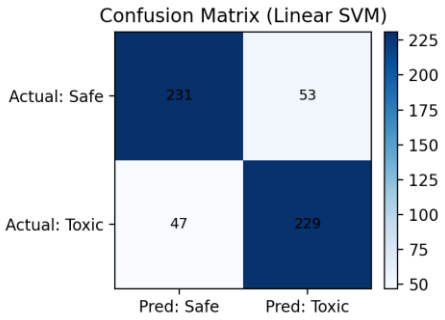

Figure 3. Confusion Matrix Visualization for Linear SVM

Figure 3 illustrates the confusion matrix of the Linear SVM model, providing an overview of correct and incorrect classifications across Safe and Toxic categories. To further examine the model's behavior at a more granular level, Table 6 presents per-class precision, recall, and F1 Scores, enabling a detailed comparison of classification performance between the two classes.

Table 6. Per-class Classification Report (Linear SVM)

Class	Precision	Recall	F1-score	Support
Safe	0,8309	0,8134	0,8221	284
Toxic	0,8121	0,8297	0,8208	276

https://doi.org/10.51276/edu.v7i2.1650

Based on Table 6, performance was balanced across classes. Safe comments reached a precision of 0.8309 and a recall of 0.8134, whereas Toxic comments reached a precision of 0.8121 and a recall of 0.8297.

Table 7. 5-Fold Cross-Validation Scores

Fold	SVM	Naïve Bayes
1	0.7714	0.7750
2	0.7607	0.7589
3	0.8339	0.8339
4	0.7352	0.7585
5	0.6369	0.6225
Mean	0.7476	0.7498
Std	±0.0642	±0.0694

Table 7 indicates variability across five folds. For SVM, accuracy ranged from 0.6369 to 0.8339 with a mean of 0.7476 (±0.0642), reflecting notable variation across partitions. To systematically understand inherent model limitations and identify potential improvement opportunities, a qualitative error analysis was conducted, examining error patterns revealed through confusion matrix decomposition. Given that the dataset fundamentally comprises informal conversational exchanges characterized by creative language use and contextual ambiguity, classification errors frequently arise from multiple interrelated sources, including contextual dependencies, semantic ambiguity, and novel obfuscation strategies not comprehensively covered in the existing slang normalization dictionary.

Table 8. Common Error Sources in Toxic Speech Classification (Qualitative)

Error Source	Description	Potential Mitigation
Sarcasm/irony	The sentence appears neutral but is intended to belittle	Expand context (thread) or use a contextual model (transformer)
Ambiguous words	Certain words can be toxic or neutral depending on context	Add bigram/trigram features and include context-aware annotation
Joking/banter	Friendly jokes between peers can resemble toxic language	Add a dedicated label or use a severity-level annotation
New obfuscation	Creative spelling to bypass filters	Expand the slang lexicon and apply character normalization
Code-mixing	Mixing Indonesian–English/Numbers	Use a multilingual model (XLM-R) or normalize code-mixed text
Very short text	One or two words provide too little context	Use a confidence threshold and human-in-the-loop review

Table 8 summarizes recurring qualitative error sources, including sarcasm/irony, context-dependent ambiguity, joking/banter, novel obfuscation patterns, code-mixing, and very short texts that provide minimal context.

https://doi.org/10.51276/edu.v7i2.1650

D. Discussion

The experiments indicate that a lightweight pipeline based on TF-IDF features and a linear Support Vector Machine can distinguish Safe and Toxic comments in Indonesian Roblox-related discourse. The approach outperformed the Multinomial Naïve Bayes baseline on the hold-out evaluation, while remaining computationally efficient and relatively transparent. These characteristics make linear classifiers attractive for an education-facing early-warning tool that is feasible to maintain in resource-constrained settings, consistent with Kusuma & Nugroho (2024), who also demonstrated the effectiveness of SVM for cyberbullying detection on Indonesian Twitter data.

Regarding RQ1, toxic utterances in this corpus are dominated by direct personal insults, performance-based mockery, and profanity, often expressed through creative spelling, abbreviations, or Indonesian-English code-mixing. These forms fit the competitive, time-pressured interaction patterns of online games and also reflect deliberate obfuscation strategies that aim to bypass simple keyword-based filtering.

Regarding RQ2, the results suggest that domain-specific slang normalization helps reduce orthographic variation and improves feature consistency, which is important in gaming communities where vocabulary changes rapidly. The variability across cross-validation folds indicates that generalization depends on the coverage of emerging domain terms. When a fold includes slang or obfuscations that are rare or absent in the training split, misclassifications increase. This implies that reliable deployment requires iterative corpus expansion and periodic lexicon updates. However, because this study did not include an ablation experiment, the isolated impact of slang normalization relative to other preprocessing steps cannot yet be quantified.

Compared to Hu et al (2025), who emphasize that cyberbullying in multiplayer games is highly contextual and shaped by competitive dynamics, the present findings provide corpus-level evidence of how these dynamics surface in Indonesian gaming talk, particularly through performance-oriented insults and stylized profanity. This supports the view that toxicity detection in gaming benefits from domain-aware linguistic handling rather than direct transfer from generic social media settings. Ismail et al (2025) similarly propose a lightweight approach for toxicity detection on gaming networks using English-language data; the present study complements this direction by demonstrating that domain-specific linguistic preprocessing can achieve practical value for Indonesian gaming communities.

Unlike Naseem et al (2025), who introduce a large gaming-chat dataset (GameTox) and baseline models primarily for English contexts, this study contributes an Indonesian-language Roblox-themed corpus and a Roblox-specific slang lexicon. In this sense, the study extends prior work by addressing a combined language and domain resource gap that remains limited in the current literature. Susanto et al (2025) further demonstrate through a multi-labeled Indonesian discourse dataset that richer annotation schemes covering toxicity, polarization, and demographics can capture nuances beyond binary classification, reinforcing the potential for expanding the present Safe–Toxic framework in future work.

https://doi.org/10.51276/edu.v7i2.1650

Studies such as Candra et al (2021) and Nabiilah et al (2023) show that transformer-based approaches can be effective for detecting Indonesian cyberbullying or toxic comments. In contrast, this work intentionally prioritizes a lighter pipeline that is easier to train, explain, and deploy with limited computational resources. For schools and families, interpretability and ease of maintenance, including updating a slang lexicon as youth language evolves, are practical advantages when the tool is positioned as a screening aid rather than an automated decision-maker.

In line with the challenges summarized by Ibrohim & Budi (2023) for Indonesian abusive-language detection, the error patterns observed here indicate that sarcasm, novel obfuscations, and code-mixed expressions remain recurring sources of mistakes, a challenge further underscored by Pamungkas & Chiril (2025), who highlight the particular difficulty of detecting hate speech in Indonesian code-mixed data. In this sense, the study From an educational standpoint, the model should be implemented as a non-punitive and human-in-the-loop early-warning mechanism. Model outputs are best treated as prompts for dialogue, reflection, and counselling rather than as evidence for sanctions. This positioning aligns with meta-analytic evidence that cyberbullying prevention is most effective when technology complements broader school and family strategies (Polanin et al., 2022; Tozzo et al., 2022) and with guidance that technology-based interventions should support educators' professional judgement (Chan et al., 2023).

E. Implication

The results of this study offer theoretical implications for research at the intersection of educational technology and Indonesian NLP. Integrating domain knowledge, such as a curated lexicon of gaming-community slang, into a standard text classification pipeline can support practical detection of harmful utterances in contexts often overlooked by general social-media studies. In addition, positioning a detection model as an educational aid reinforces the principle that algorithmic outputs should be interpreted within pedagogical frameworks rather than treated as standalone judgements.

In practical terms, educators and school counsellors may use the model as a preliminary screening tool to help map potentially harmful communication patterns that students encounter in gaming-related digital spaces. When used responsibly, the predicted outputs can serve as concrete discussion materials in digital literacy and character education activities, supporting classroom dialogue about empathy, respectful communication, and the consequences of online speech. The tool also helps counsellors prioritize follow-up conversations by highlighting texts that warrant closer human review, while keeping final decisions with trained educators.

For parents and families, the slang lexicon embedded in the pipeline can serve as a learning resource to understand better gaming terminology and how children express emotion online. This understanding can enable more constructive parental mediation and earlier conversations about self-regulation and digital safety. At the school-management level, aggregated and anonymized trends, if collected ethically, may inform child-friendly

https://doi.org/10.51276/edu.v7i2.1650

school policies, anti-bullying campaigns, and the design of preventive programs that extend beyond offline settings into digital interactions commonly experienced by students.

The Streamlit prototype can also be leveraged as educational media to demonstrate how text classification works and to promote critical AI literacy among teachers, students, and parents. To avoid misuse, implementation should prioritize privacy protection, transparency, and a strict human-in-the-loop workflow. In particular, predictions should not be used for automatic punishment; instead, they should be treated as early signals to support guidance, counselling, and restorative educational interventions.

F. Limitation and Suggestion for Further Research

This investigation acknowledges several methodological limitations that constrain the interpretation and generalization of findings. First, the data corpus originates exclusively from external online communities discussing Roblox gaming content rather than from in-game chat, potentially limiting generalizability to authentic school contexts and in-game interaction patterns where communication dynamics may differ substantially. Second, the classification framework employs a binary Safe versus Toxic distinction, thereby failing to capture the nuanced spectrum of toxicity types that could inform more targeted educational interventions, such as distinguishing between personal insults, competitive trash talk, discriminatory harassment, and explicit threats. Third, the current model architecture does not incorporate conversational thread context or temporal sequencing information, missing potentially valuable signals about escalation patterns and relationship dynamics that could enhance detection accuracy and enable more sophisticated intervention strategies.

Fourth, annotation reliability assessment using inter-rater agreement was not conducted, given the single-annotator approach used for practical feasibility, potentially introducing subjective bias that could affect label quality and model training. Fifth, and critically from an educational perspective, this investigation evaluates model performance characteristics rather than measuring the actual effectiveness of educational interventions, leaving empirical evidence of real-world impact on student behavior, school climate, or educational outcomes absent. Future research should address these limitations through pilot studies in authentic educational settings to assess practical utility and measure educational impact outcomes.

Recommendations for subsequent research investigations encompass multiple dimensions. First, systematic ablation studies should be conducted to precisely quantify the specific contribution of slang normalization procedures to overall classification performance, enabling evidence-based optimization of preprocessing pipelines. Second, data collection efforts should be systematically expanded across diverse temporal periods to ensure comprehensive coverage of emerging slang expressions and linguistic innovations that characterize rapidly evolving youth gaming communities. Third, classification frameworks should be extended to multi-label taxonomies, distinguishing specific toxicity categories such as insults, threats, harassment, and profanity, potentially incorporating

https://doi.org/10.51276/edu.v7i2.1650

severity ratings that inform proportionate intervention strategies, as demonstrated by Findawati et al (2025), who show that keyword-driven ensemble classifiers can effectively handle multi-label dangerous speech classification on imbalanced data. Fourth, architectural enhancements should incorporate conversational context through sequential modeling approaches that capture thread-level dynamics and escalation patterns.

Fifth, controlled pilot studies should be conducted within educational institutional settings to systematically evaluate the practical utility of the tool as a digital literacy educational tool and a school counseling support resource, measuring both technical performance in authentic deployment conditions and educational outcome effectiveness through student surveys, educator interviews, and behavioral incident tracking. Sixth, comprehensive bias and fairness evaluations should be conducted to examine potential disparities across linguistic variations, demographic groups, and cultural contexts, ensuring equitable and just application across diverse student populations. These research directions collectively advance both technical capabilities and educational implementation readiness of automated toxicity detection systems.

G. Conclusion

This investigation successfully constructed and systematically evaluated a comprehensive detection pipeline for identifying cyberbullying-related toxic speech within Indonesian-language comments themed around Roblox gaming communities. Employing TF-IDF unigram-bigram vectorization combined with Linear Support Vector Machine classification, the implemented system achieved an accuracy of 82.14% and a macro-averaged F1-score of 0.8214 on hold-out testing comprising 560 samples, demonstrating a clear advantage over the Naïve Bayes baseline across all evaluated metrics.

Regarding research implications, these empirical findings demonstrate that computationally efficient lightweight model approaches that maintain interpretability and enable rapid prototyping through frameworks such as Streamlit constitute valuable instrumental support mechanisms for digital character education initiatives and digital literacy development programs. The system functions optimally as a non-punitive screening tool designed to assist school counselors and parents in conducting more informed educational interventions and implementing earlier prevention measures while strictly maintaining human-in-the-loop decision-making principles that preserve essential pedagogical judgement and relationship-based intervention capabilities.

Future research investigations should prioritize conducting controlled field studies within authentic educational institutional environments, systematically expanding data collection across diverse temporal periods to capture linguistic evolution, implementing rigorous ablation studies to isolate the specific contribution of slang normalization procedures, and developing enhanced multi-label classification frameworks incorporating conversational context modeling to enable more accurate detection and more appropriately calibrated educational intervention responses. These research directions would advance

https://doi.org/10.51276/edu.v7i2.1650

both technical detection capabilities and the effectiveness of educational implementation in supporting healthy digital socialization among young people.

References

Balakrisnan, V., & Kaity, M. (2023). Cyberbullying Detection and Machine Learning: A Systematic Literature Review. *Artificial Intelligence Review*, *56*(Suppl 1), 1375–1416. https://doi.org/10.1007/s10462-023-10553-w

Bustamin, A., Prayogi, A. A., Siswanto, D., Rafrin, M., & Nurdin, A. (2025). Text Normalization for Indonesian Slang Words in Sentiment Analysis Development. *ICIC Express Letters, Part B: Applications*, *16*(2), 121–129. https://doi.org/10.24507/icicelb.16.02.121

Candra, A., Wella, & Wicaksana, A. (2021). Bidirectional Encoder Representations from Transformers for Cyberbullying Text Detection in Indonesian Social Media. *International Journal of Innovative Computing, Information and Control*, *17*(5), 1599–1615. https://doi.org/10.24507/ijicic.17.05.1599

Chan, N. N., Samsudin, N., Hoo, M. C., Ridzuan, M. I. B. M., Ooi, P. B., Mohamad, A. M. A. M., & Scheithauer, H. (2023). The Digital Defence against Cyberbullying: A Systematic Review of Tech-Based Approaches. *Cogent Education*, *10*(2), 2288492. https://doi.org/10.1080/2331186X.2023.2288492

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Elisabeth, D., Budi, I., & Ibrohim, M. O. (2020). Hate Code Detection in Indonesian Tweets Using Machine Learning Approach: A Dataset and Preliminary Study. In *2020 8th International Conference on Information and Communication Technology (ICoICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICoICT49345.2020.9166251

Findawati, Y., Raharjo, A. B., Navastara, D. A., Yonathan, V., Yatestha, A. A., & Purwitasari, D. (2025). Multi-Label Aspect Dangerous Speech Classification Using Keyword-Driven Ensemble Classifier on Imbalanced Data. *JOIV: International Journal on Informatics Visualization*, *9*(4), 3129. https://doi.org/10.62527/joiv.9.4.3129

Hibatullah, H., Ballı, T., & Yetkin, E. F. (2025). Verbal Harassment Detection in Online Games Using Machine Learning Methods. *Entertainment Computing*, *55*, 101009. https://doi.org/10.1016/j.entcom.2025.101009

Hu, Y., Clancy, E. M., & Klettke, B. (2025). Player versus Player: A Systematic Review of Cyberbullying in Multiplayer Online Games. *Computers in Human Behavior Reports*, *18*, 100675. https://doi.org/10.1016/j.chbr.2025.100675

Ibrohim, M. O., & Budi, I. (2023). Hate Speech and Abusive Language Detection in Indonesian Social Media: Progress and Challenges. *Heliyon*, *9*(8), e18647. https://doi.org/10.1016/j.heliyon.2023.e18647

https://doi.org/10.51276/edu.v7i2.1650

Ismail, M., Jones, B. C., & Fadzil, A. F. (2025). Enhancing Online Toxicity Detection on Gaming Networks: A Multi-Feature, Lightweight Approach. *Crime Science*, *14*(1), 2. https://doi.org/10.1007/s41060-025-00730-1 ⚠

Isnawan, F. (2025). Pencegahan Cyberbullying melalui Pendidikan Karakter dan Pendidikan Hukum bagi Siswa Sekolah. *Jurnal Civic Hukum*, *10*(1). https://doi.org/10.22219/jch.v10i1.36879

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning (ECML 1998)* (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Kusuma, R., & Nugroho, A. (2024). Deteksi Cyberbullying pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM). *JUTIF: Jurnal Teknik Informatika*, *5*(1). https://doi.org/10.52436/1.jutif.2024.5.1.809

Marciano, L., Schulz, P. J., & Camerini, A.-L. (2020). Cyberbullying Perpetration and Victimization in Youth: A Meta-Analysis of Longitudinal Studies. *Journal of Computer-Mediated Communication*, *25*(2), 163–181. https://doi.org/10.1093/jcmc/zmz031

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT Base Model for Toxic Comment Analysis on Indonesian Social Media. *Procedia Computer Science*, *216*, 714–721. https://doi.org/10.1016/j.procs.2022.12.188

Naseem, U., Shiwakoti, S., Shah, S. B., Thapa, S., & Zhang, Q. (2025). GameTox: A Comprehensive Dataset and Analysis for Enhanced Toxicity Detection in Online Gaming Communities. In *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 440–447). https://doi.org/10.18653/v1/2025.naacl-short.37

Pamungkas, E. W., & Chiril, P. (2025). Ngalawan Ujaran Sengit: Hate Speech Detection in Indonesian Code-Mixed Social Media Data. *Language Resources and Evaluation*, *59*, 2387–2414. https://doi.org/10.1007/s10579-025-09810-x

Polanin, J. R., Espelage, D. L., & Grotpeter, J. K. (2022). A Systematic Review and Meta-Analysis of Interventions to Decrease Cyberbullying Perpetration and Victimization. *Prevention Science*, *23*(3), 439–454. https://doi.org/10.1007/s11121-021-01259-y

Putri, S. D. A., Ibrohim, M. O., & Budi, I. (2021). Abusive Language and Hate Speech Detection for Javanese and Sundanese Languages in Tweets: Dataset and Preliminary Study. In *Proceedings of the 2021 International Conference on World Computing and Software Engineering (WCSE)* (pp. 67–72). https://doi.org/10.18178/wcse.2021.02.011

Susanto, L., Wijanarko, M. I., Pratama, P. A., Tang, Z., Akyas, F., Hong, T., Idris, I. K., Aji, A. F., & Wijaya, D. T. (2025). A Multi-Labeled Dataset for Indonesian Discourse:

https://doi.org/10.51276/edu.v7i2.1650

Examining Toxicity, Polarization, and Demographics Information. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 18863–18890). https://doi.org/10.18653/v1/2025.findings-acl.966

Tozzo, P., Cuman, O., Moratto, E., & Caenazzo, L. (2022). Family and Educational Strategies for Cyberbullying Prevention: A Systematic Review. *International Journal of Environmental Research and Public Health*, *19*(16), 10452. https://doi.org/10.3390/ijerph191610452

Author's Biography

	**Farhan Badrani.** He was born in Purwakarta on 26 September 2003. He is a student in the Bachelor's program of Information Systems and Technology Education at Universitas Pendidikan Indonesia (UPI), Purwakarta Campus, Class of 2021. Email: farhanbadrani23@upi.edu
	**Nuur Wachid Abdul Majid.** He was born in Kulon Progo on 25 June 1991. He is a Lecturer and Head of the Study Program of Information Systems and Technology Education (Pendidikan Sistem dan Teknologi Informasi) at Universitas Pendidikan Indonesia (UPI), Purwakarta Campus. Email: nuurwachid@upi.edu