Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Cyberbullying Detection in Indonesian TikTok Comments Using IndoBERT with Fairness Evaluation

Hanik Dewi Jayanti[1], Abdul Rohman[2]

[1,2] Informatics and Computer Engineering Education Program, Univeristas Ngudi Waluyo, Indonesia

Received:
December 14, 2025
Revised:
January 23, 2026
Accepted:
February 2, 2026
Published:
March 2, 2026

Corresponding Author:
Author Name*:
Hanik Dewi Jayanti
Email*:
hanikdewi004@gmail.com

DOI:
10.63158/journalisi.v8i1.1448

Abstract. This study investigates automated cyberbullying detection on TikTok within the Indonesian digital context, where high social media usage among children and adolescents demands scalable and consistent content moderation. We propose an IndoBERT-based framework for detecting and classifying cyberbullying in Indonesian-language TikTok comments, incorporating algorithmic fairness considerations. A dataset of 2,122 TikTok comments was collected from a publicly available Kaggle repository and divided into training, validation, and testing sets using a 70:15:15 stratified sampling ratio. The IndoBERT-base-p1 model was fine-tuned with the PyTorch and HuggingFace frameworks, optimizing hyperparameters like the AdamW optimizer and learning rate scheduling. Experimental results show that the model achieved an accuracy of 70.66% and a ROC-AUC score of 0.7969, demonstrating solid discriminative power. With a macro F1-score of 0.7066 and a cyberbullying recall of 0.7170, the model shows balanced performance in identifying harmful content. A key contribution of this study is a fairness evaluation framework that reveals an accuracy gap of 2.08% and an equal opportunity gap of 0.0208, indicating overall fairness. However, demographic parity remains a concern. This system, supporting content triage combined with human review, enhances moderation workflows by filtering non-cyberbullying cases while flagging potentially harmful content for human oversight.

Keywords: Content Moderation, Cyberbullying Detection, Fairness Evaluation, IndoBERT, TikTok Comments

1. INTRODUCTION

The rapid evolution of information and communication technologies has positioned social media as a fundamental communication infrastructure, facilitating cross-geographical information exchange and reshaping social behavior and learning patterns across multigenerational societies. However, these platforms also introduce significant challenges in ensuring the safety of vulnerable groups, particularly children and adolescents [1]. Social media platforms, notably TikTok, have gained immense popularity among young users worldwide, not only as sources of entertainment but also as important spaces for mental health discourse and public communication [2]. In Indonesia, the digital ecosystem is expanding rapidly, with around 64 million internet users and 10 million TikTok users, and young generations spending an average of nine hours per day online. This creates an environment where social media, especially short-video platforms, can both foster creativity and facilitate harmful behaviors such as cyberbullying, highlighting the need for automated detection and classification systems to address these issues [3].

Despite the potential of manual moderation to identify harmful content, it is limited by scalability, consistency, and objectivity, especially given the sheer volume of social media content. As a result, the adoption of automated approaches using Natural Language Processing (NLP) techniques is becoming increasingly essential to efficiently detect and classify cyberbullying. Traditional machine learning classifiers like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes have been applied to this task, achieving accuracy rates exceeding 90%. However, these methods often rely on N-gram and TF-IDF features and classification schemes that may not be suitable for informal, slang-heavy online discourse [4]. The rapid growth of social media platforms and their increasing use by vulnerable groups such as children and adolescents intensifies the urgency of implementing more effective detection mechanisms.

The field of cyberbullying detection has seen a shift towards more advanced neural network architectures, such as Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNNs), with some studies reporting performance improvements over conventional methods. More recently, Transformer-based

architectures like BERT have shown superior performance due to their ability to capture bidirectional contextual representations. However, the majority of these studies have been conducted in high-resource languages, leaving a significant gap in research on the application of such models to low-resource languages like Indonesian. Indonesian presents unique linguistic challenges, including informal expressions, internet slang, and code-switching, which are rarely addressed by existing detection frameworks [6].

Despite the promising advancements in cyberbullying detection, several key gaps remain in the context of Indonesian social media. First, while Transformer-based models such as BERT have demonstrated superior performance in high-resource languages, their application to the Indonesian language—especially in informal TikTok comments laden with slang—has not been sufficiently explored [7]. Second, existing studies have focused largely on model accuracy, often overlooking fairness considerations, which are crucial for ensuring responsible AI deployment in real-world content moderation systems [8]. Third, the linguistic nuances of Indonesian, including code-switching, regional dialects, and evolving internet slang, pose distinct challenges that current models do not adequately address [2][9]. These gaps highlight the need for tailored solutions that consider both the linguistic and fairness aspects of cyberbullying detection.

This study aims to address these gaps through three primary contributions. First, we fine-tune IndoBERT, a language-specific pre-trained model, on a dataset of Indonesian TikTok comments that includes authentic slang expressions, improving the detection of cyberbullying in informal digital communication. Second, we propose a comprehensive fairness analysis framework, incorporating metrics such as Accuracy Gap, Demographic Parity, and Equal Opportunity, to systematically evaluate algorithmic fairness. Finally, we conduct an in-depth error analysis and performance evaluation based on comment length, identifying specific limitations of the model and providing actionable insights for future improvements. These contributions not only advance the theoretical understanding of fairness-aware NLP models for low-resource languages but also support the development of responsible content moderation systems tailored to the unique needs of Indonesian social media platforms.

2. METHODS

2.1. Data Preparation and Exploration

This study utilizes a dataset of Indonesian-language TikTok comments obtained from the Kaggle repository entitled "Cyberbullying Dataset with Slang" [10], which is distributed under the CC0: Public Domain license, permitting unrestricted use for research purposes without requiring attribution. The dataset was selected based on several strategic considerations. First, it specifically contains comments collected from the TikTok platform, which constitutes the primary focus of this research. Second, the dataset incorporates a wide range of slang expressions, reflecting a distinctive characteristic of social media communication in the Indonesian context. Third, the dataset has undergone an initial curation process, thereby facilitating subsequent stages of data preprocessing and analysis.

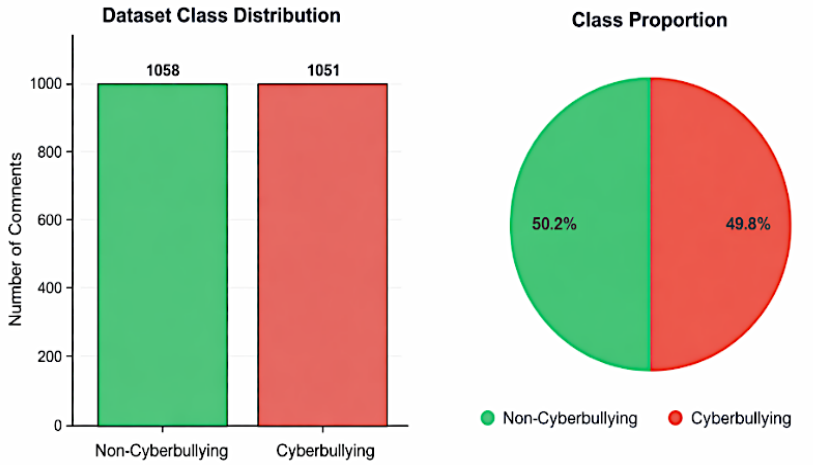

Figure 1. Distribution of Datasets

The dataset consists of 2,122 TikTok comment samples categorized into two primary labels: Non-Cyberbullying and Cyberbullying. The class distribution is illustrated in Figure 1, which depicts the proportion of each category to provide an overview of the dataset balance. Each sample in the dataset is structured to include a textual comment attribute and an annotated category label. These characteristics reflect authentic communication patterns among Indonesian TikTok users and present inherent challenges for automated cyberbullying detection. Furthermore, the use of TikTok comment data in this study adheres to established research ethics principles, including

the anonymization of user-identifiable information and the utilization of publicly available data in compliance with Indonesia's personal data protection regulations, specifically Law Number 27 of 2022 on Personal Data Protection [11].

1) Dataset Preprocessing

The preprocessing stage is defined as a series of systematic transformation protocols applied to raw textual data with the objective of producing a standardized corpus that can be optimally processed within machine learning model architectures for sentiment analysis and content classification tasks [12]. According to Aliyah et al., prior studies have demonstrated that the removal of non-informative elements combined with text normalization can significantly enhance a model's ability to identify meaningful linguistic patterns [9]. The preprocessing pipeline was implemented as a Python function applied to all comments prior to dataset splitting, ensuring consistency across training, validation, and test sets. Importantly, we retained slang terms, informal abbreviations, and code-switching patterns without normalization, as these are characteristic features of cyberbullying discourse on Indonesian TikTok and critical for realistic model evaluation.

2) Dataset Splitting

A stratified split ensures that each class is proportionally represented across all dataset partitions, maintaining the original class distribution in training, validation, and test sets [13]. The 70:15:15 data split ratio adopted in this study follows common practices in deep learning research for text classification tasks, balancing the need for sufficient training data with adequate validation and independent test sets. [14]. Stratified sampling was implemented using the train_test_split function from the scikit-learn library (version 1.3.0), with the stratify parameter set to the class labels [15]. Given the total dataset size of 2,122 samples with 1,062 Non-Cyberbullying and 1,060 Cyberbullying comments, the splitting procedure was conducted in two stages:

a) First Stage: The dataset was split into training set (70%, n=1,485) and a combined validation-test set (30%, n=637), with random_state=42 for reproducibility.
b) Second Stage: The combined validation-test set was further divided equally into validation set (15% of total, n=318) and test set (15% of total, n=319), again using stratification and the same random seed.

c) The final dataset distribution is as follows: Training set: 1,485 samples (743 Non-Cyberbullying, 742 Cyberbullying), Validation set: 318 samples (159 Non-Cyberbullying, 159 Cyberbullying), and Test set: 319 samples (160 Non-Cyberbullying, 159 Cyberbullying).

d) This stratified approach ensures proportional class representation (approximately 50:50) in all partitions while maintaining statistical independence between sets for unbiased model evaluation.

2.2. Model Development Using IndoBERT

This study implements the IndoBERT model using the PyTorch framework, which provides imperative programming flexibility for efficient deep learning architecture development [16], and employs HuggingFace as a platform to access and fine-tune pre-trained IndoBERT models for Indonesian-language cyberbullying classification tasks [17]. IndoBERT is a pre-trained language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically trained on large-scale Indonesian corpora, making it more effective in capturing the nuances and linguistic structures of the Indonesian language compared to multilingual BERT or BERT models trained on other languages [7]. In this study, IndoBERT-base-p1 is adopted as the model backbone, featuring a configuration of 12 transformer layers with 110 million parameters trained on a massive corpus of approximately 220 million Indonesian-language tokens from diverse domains, including social media. This extensive pre-training enables the model to effectively capture informal linguistic characteristics prevalent in TikTok comments for cyberbullying classification tasks [18]. The IndoBERT architecture is augmented with 128 trainable prompt tokens (also called pseudo-tokens or soft prompts) prepended to the input sequence, following the prompt tuning methodology [19]. These learnable embeddings serve as task-specific context adapters, enabling the model to extract semantic representations optimized for cyberbullying detection without modifying the pre-trained transformer weights. This approach provides parameter-efficient fine-tuning while maintaining the model's ability to generalize across diverse input lengths and syntactic structures.

The optimization of IndoBERT model performance in this study is achieved through a systematic hyperparameter tuning process aimed at identifying optimal parameter configurations such as learning rate, batch size, and number of training epochs to

maximize cyberbullying detection accuracy while minimizing classification bias. External configuration parameters, including learning rate and batch size, which require substantial computational resources, are carefully tuned, whereas internal components such as weight matrices and bias vectors are adaptively updated during the training phase using the labeled TikTok comment corpus [20]. The hyperparameter configuration employed during the training and fine-tuning of the IndoBERT model is presented in Table 1.

Table 1. Hyperparameter Configuration

Parameters	Value	Description
Max Sequence Length (MAX_LEN)	128	Maximum length of text input tokens.
Batch Size	16	Number of samples per training iteration.
Number of Epochs	3	Number of model training cycles.
Learning Rate	2e-5	Weight update rate during fine-tuning.
Warmup Ratio	0.1	Initial step size for gradually increasing the learning rate.

The optimization of the IndoBERT model in this study employs the AdamW algorithm, which decouples learning rate regulation from weight decay, thereby applying weight regularization directly at each iteration without modifying the loss function. This approach provides superior convergence stability when handling complex cyberbullying datasets [21].

2.3. Training and Validation Process

The training loop is conducted through forward and backward propagation using gradient descent to minimize the loss function, accompanied by periodic validation as a feedback mechanism for model optimization [22]. During the backward pass, gradient clipping is applied to prevent exploding gradients, after which parameter updates are sequentially executed using optimizer step followed by scheduler step. Subsequently, the validation process is performed at the end of each epoch in evaluation mode with torch no_grad to enhance memory efficiency. Model checkpoints automatically save the state_dict corresponding to the highest validation accuracy into a file named best model. Additionally, the training history including metrics such as training accuracy

(train_acc), training loss (train_loss), validation accuracy (val_acc), and validation loss (val_loss) is recorded in training_history csv and visualized through accuracy and loss curves to facilitate convergence analysis.

2.4. Model Evaluation

Model evaluation is conducted using the test set, which comprises 15% of the total dataset and is randomly separated prior to the commencement of the training process to ensure evaluation independence [15]. The evaluation metrics include overall performance measures, specifically Accuracy, which is computed using Equation 1.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the classification outcomes. Precision and Recall are computed for each class using Equations 2 and 3, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

Meanwhile, the F1-score, defined as the harmonic mean of Precision and Recall, is formulated in Equation 4 [23].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The ROC–AUC score is calculated by integrating the ROC curve, as shown in Equation 5.

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)) \quad 5$$

where the True Positive Rate (TPR), also referred to as sensitivity, and the False Positive Rate (FPR) are defined in Equation 6.

$$TPR=\frac{TP}{TP+FN},FPR=\frac{FP}{FP+TN} \tag{6}$$

The prediction process produces y_pred (predicted labels), y_pred_probs (class probabilities), and y_true (ground truth labels), which are stored in the test_predictions.csv file. The confusion matrix is visualized in two formats: raw counts displayed as an absolute heatmap and a normalized version expressed as row-wise percentages to facilitate interpretation in imbalanced datasets. The normalized confusion matrix is computed using Equation 7.

$$C_{norm}[i,j]=\frac{C[i,j]}{\sum_j C[i,j]} \tag{7}$$

Receiver Operating Characteristic (ROC) analysis explores the trade-off between sensitivity and specificity across various decision thresholds without bias arising from arbitrary cutoff selection. The aggregated ROC–AUC metric quantifies the model's discriminative capability on a scale ranging from 0 to 1, where curves converging toward the upper-left corner indicate optimal classification performance in cyberbullying detection [24]. The evaluation results are visualized through a metrics comparison presented as a bar chart for five key metrics, featuring color-coded bars, value labels, and a baseline reference line (0.5). The classification report is stored as a text file containing class-wise precision, recall, and F1-score, along with support, macro-average, and weighted-average values. All evaluation metrics are recorded in the evaluation_metrics.csv file, and all visualizations are saved in PNG format with a resolution of 300 dpi for documentation and analytical purposes.

Additionally, we compute the Matthews Correlation Coefficient (MCC), which is particularly robust for binary classification tasks with imbalanced datasets, as it accounts for all four confusion matrix categories (TP, TN, FP, FN) and returns values between -1 and +1 [24]. MCC is calculated using Equation 8.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

An MCC value of +1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between predictions and observations. MCC is especially valuable for our cyberbullying detection task as it provides a balanced performance measure that is not inflated by class imbalance, unlike accuracy which can be misleading when classes are nearly balanced but model performance differs significantly across classes.

2.5. Class-wise Error Balance and Fairness Analysis

1) Fairness Metrics Framework

Fairness analysis is defined as the evaluation of error balance and classification equity of the IndoBERT model across Cyberbullying and Non-Cyberbullying classes, focusing on the symmetry of false positive and false negative rates rather than demographic parity [25]. Since our dataset does not contain demographic metadata (e.g., user age, gender, ethnicity), we reframe traditional fairness metrics to assess class-wise error balance a critical consideration for content moderation systems where both over-censorship (false positives) and under-detection (false negatives) carry significant consequences.

This evaluation is conducted using three complementary metrics: Accuracy Gap, Demographic Parity (interpreted as prediction rate balance), and Equal Opportunity (measuring recall parity). While demographic parity is traditionally used to assess fairness across protected demographic groups, in this study it serves as a proxy for evaluating whether the model exhibits systematic bias toward predicting one class more frequently than the other, independent of ground truth labels. This reinterpretation is appropriate given our class-based (rather than demographic group-based) fairness evaluation framework.

2) Error Analysis

Error analysis is conducted systematically to identify, categorize, and understand the prediction errors made by the IndoBERT model, serving as an essential diagnostic procedure to uncover architectural limitations and provide insights for model

improvement [26]. Prediction errors are categorized into two main types based on the confusion matrix: False Positives (FP), where *Non-Cyberbullying* comments are incorrectly classified as *Cyberbullying*, and False Negatives (FN), where *Cyberbullying* comments fail to be detected by the model.

3) Evaluation and Success Criteria

The model is considered to satisfy algorithmic fairness criteria if: (1) the Accuracy Gap is less than 0.05, (2) the Demographic Parity Gap is less than 0.1, and (3) the Equal Opportunity Gap is less than 0.1, as presented in Table 2. All metrics are evaluated against the predefined thresholds using a recommendation logic to provide guidance for model improvement based on the metrics that fail to meet the established standards.

Table 1. Fairness and Threshold Metrics

Metric	Formula	Threshold
Accuracy Gap	$\lvert Acc_0 - Acc_1 \rvert$	< 0.05
Demographic Parity Gap	$\lvert PPR_0 - PPR_1 \rvert$	< 0.1
Equal Opportunity Gap	$\lvert TPR_0 - TPR_1 \rvert$	< 0.1

Information: 0 = Non-Cyberbullying, 1 = Cyberbullying

3. RESULTS AND DISCUSSION

This section presents the experimental results and discussion in an integrated manner to address the research problems related to the detection and classification of cyberbullying in Indonesian-language TikTok comments using IndoBERT. The presentation of results focuses on model performance evaluation and fairness analysis, while the discussion emphasizes the implications, contributions, and comparisons with findings from previous studies.

3.1. Baseline Models Comparison

To establish the effectiveness of the IndoBERT model, we first evaluated two baseline approaches commonly used for text classification tasks:

1) TF-IDF with Logistic Regression

Text was vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) with a maximum of 5,000 features, and classified using Logistic Regression with L2 regularization (C=1.0). This baseline achieved an accuracy of 64.26%, precision of 0.63, recall of 0.67, and F1-score of 0.65.

2) TF-IDF with Support Vector Machine (SVM)

Using the same TF-IDF features, we rained a linear SVM classifier with C=1.0, which achieved an accuracy of 66.15%, precision of 0.65, recall of 0.69, and F1-score of 0.67.

Table 3. presents a comprehensive comparison between baseline models and the IndoBERT approach.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
TF-IDF + Logistic Regression	0.6426	0.63	0.67	0.65	0.69
TF-IDF + SVM	0.6615	0.65	0.69	0.67	0.71
IndoBERT (Ours)	0.7066	0.70	0.72	0.71	0.7969
Improvement over SVM	+6.81%	+7.69%	+4.35%	+5.97%	+12.24%

The results demonstrate that IndoBERT substantially outperforms both traditional machine learning baselines, with an accuracy improvement of 6.4 percentage points over TF-IDF+SVM and 4.5 points in F1-score. This performance gain can be attributed to IndoBERT's ability to capture contextual semantics, handle informal language patterns, and leverage pre-trained knowledge of Indonesian linguistic structures capabilities that TF-IDF representations cannot provide. The superior ROC-AUC score of IndoBERT (0.7969 vs. 0.71 for SVM) further confirms its stronger discriminative power in separating cyberbullying from non-cyberbullying content.

3.2. Model Performance Evaluation

Model performance evaluation was conducted on a test dataset consisting of 317 TikTok comments. The experimental results indicate that the IndoBERT model achieved an overall accuracy of 70.66%. Table 3 presents a summary of the primary evaluation metrics obtained from the model testing process.

Table 4. Model Performance Metrics

Metric	Value
Accuracy	0.7066
Precision (Macro)	0.7067
Recall (Macro)	0.7066
F1-Score (Macro)	0.7066
ROC-AUC	0.7969

The relatively balanced macro- and weighted-average F1-scores indicate that the model performs consistently across both classes, namely cyberbullying and non-cyberbullying. A ROC–AUC value of 0.7969 demonstrates that the model exhibits good discriminative capability in distinguishing between comments that contain cyberbullying and those that do not. The ROC curve visualization is presented in Figure 2 to further illustrate the model's ability to separate the two classes across different decision threshold values.

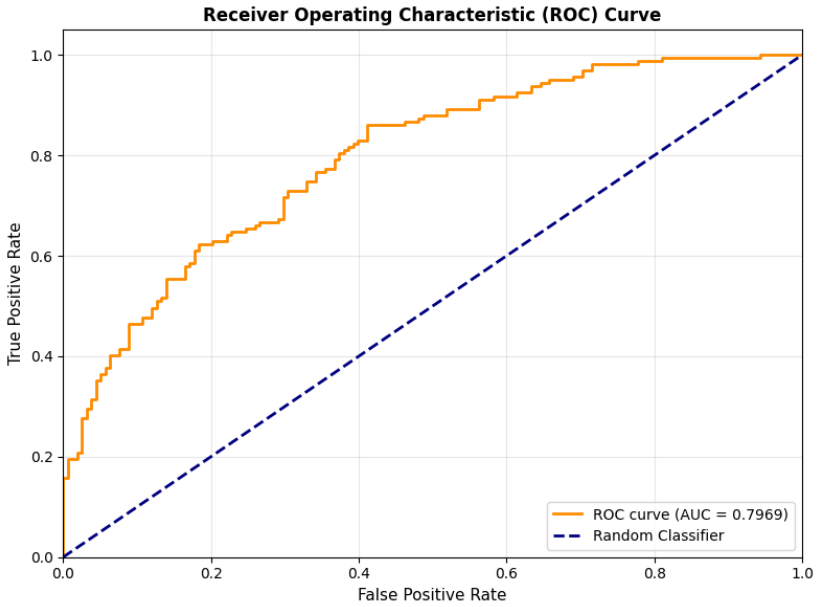

Figure 2. Receiver Operating Characteristic (ROC) Curve of IndoBERT Cyberbullying Classification Model

Based on the per-class evaluation, the model demonstrates relatively balanced performance. The cyberbullying class achieves a recall value of 0.7170, indicating that the model is sufficiently sensitive in detecting problematic comments. This level of sensitivity is particularly important in the context of social media content moderation, as failure to detect cyberbullying may lead to adverse consequences for users. The

confusion matrix obtained from the model evaluation is presented in Table 4, illustrating the distribution of correct and incorrect predictions for each class.

Table 5. Confusion Matrix

Actual \ Prediction	Cyberbullying	Non-Cyberbullying
Cyberbullying	114	45
Non-Cyberbullying	48	110

To facilitate visual interpretation of the classification errors, the confusion matrix results are also visualized in Figure 3.

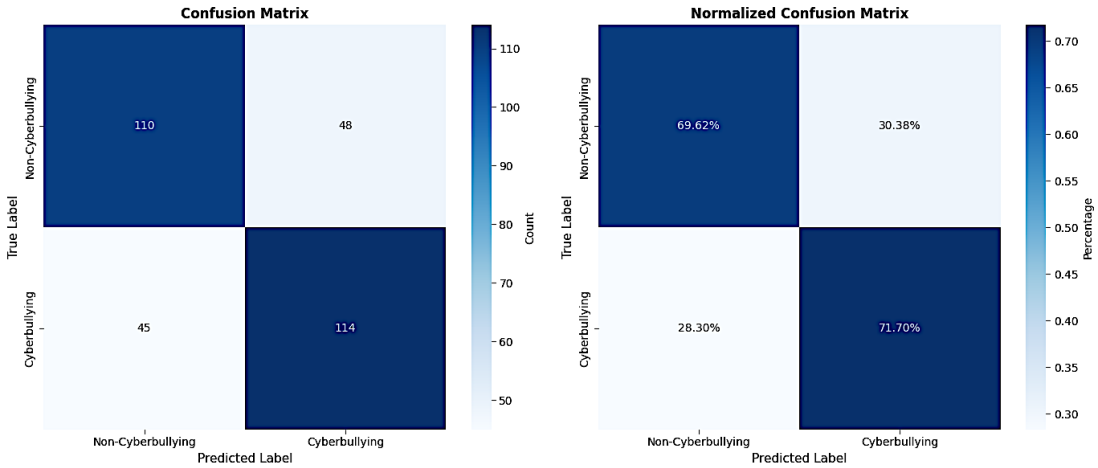

Figure 3. Confusion Matrix Visualization of Cyberbullying Classification Using IndoBERT

From the confusion matrix, it can be observed that the model correctly classified 114 cyberbullying comments out of a total of 159 (true positives), while 45 cyberbullying comments were incorrectly classified as non-cyberbullying (false negatives). For the non-cyberbullying class, the model correctly identified 110 comments out of 158 (true negatives), whereas 48 non-cyberbullying comments were misclassified as cyberbullying (false positives).

3.3. Model Fairness Analysis

In addition to performance evaluation, this study also conducts a fairness analysis to identify potential model bias in predicting the cyberbullying and non-cyberbullying

classes. This analysis is essential to ensure that the model is not only accurate but also fair when deployed in content moderation systems. The results indicate that the accuracy for the non-cyberbullying class is 0.6962, while the accuracy for the cyberbullying class is 0.7170. The accuracy gap between classes is 2.08%, suggesting that the model demonstrates a good level of fairness in terms of performance, as this gap falls below the commonly used 5% threshold for indicating performance unfairness. A summary of the fairness metrics is presented in Table 5.

Table 2. Fairness Metrics Summary

Metric	Value
Accuracy (Non-Cyberbullying)	0.6962
Accuracy (Cyberbullying)	0.7170
Accuracy Gap	0.0208
False Positive Rate	0.3038
False Negative Rate	0.2830
Demographic Parity Gap	0.4132
Equal Opportunity Gap	0.0208

The low Equal Opportunity Gap value (0.0208) indicates that the model provides nearly equal opportunities for correctly identifying *cyberbullying* and *non-cyberbullying* comments. This balanced performance across classes is visualized in Figure 4, which presents a comparison of per-class accuracy and prediction error rates.

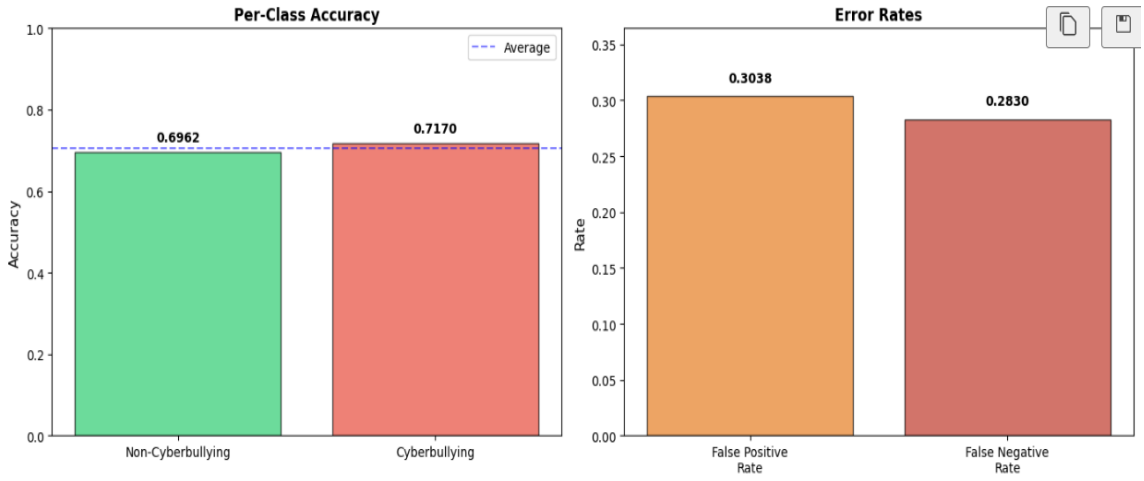

Figure 4. Accuracy and Error Rate per Class of the Cyberbullying Classification Model

This indicates that the model does not significantly disadvantage either class during the classification process. These findings strengthen the argument that IndoBERT can serve as a relatively fair foundation for a *cyberbullying* detection system. However, the relatively high Demographic Parity Gap value (0.4132) suggests a substantial difference in positive prediction rates across classes. The distribution of the model's prediction confidence for each class is illustrated in Figure 5, showing that *cyberbullying* comments tend to receive higher predicted probabilities compared to *non-cyberbullying* comments. This phenomenon is commonly observed in content-based text classification tasks, as the linguistic characteristics of cyberbullying comments tend to be more explicit and context-rich. Therefore, the demographic parity metric should be interpreted with caution and should not be used as the sole indicator of fairness.

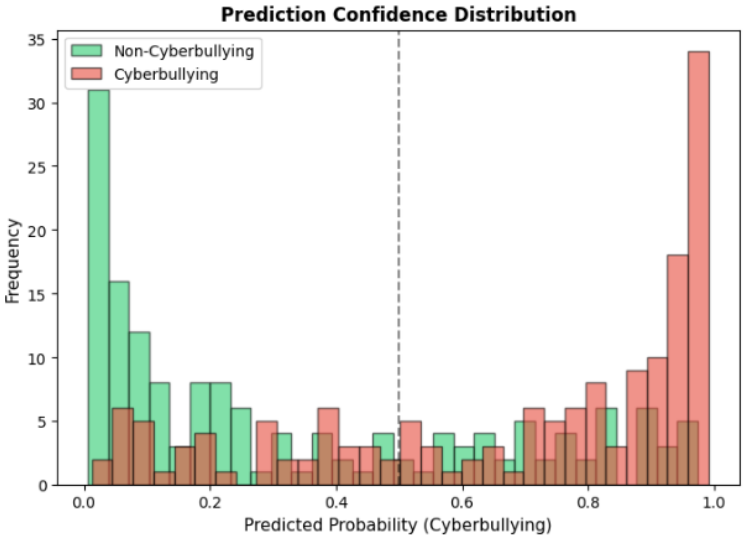

Figure 5. Confidence Distribution Predictions for Cyberbullying and Non-Cyberbullying Classes

3.4. Effect of Comment Length on Model Performance

Additional analysis based on comment length indicates that model accuracy increases as text length grows, with the highest accuracy observed in the longest comment quartile. The comparison of model performance across comment length quartiles is

visualized in Figure 6, providing a clearer illustration of the effect of text length on model performance.

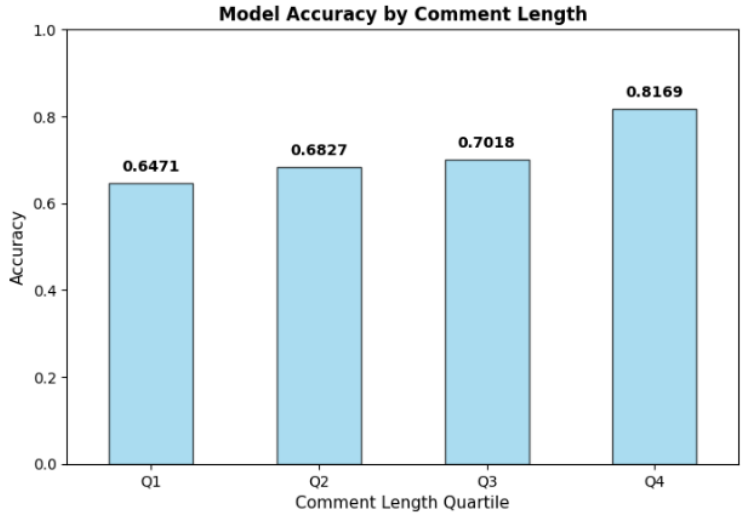

Figure 6. Model Accuracy Based on Comment Length Quartiles

The distribution of correct and incorrect predictions across each comment-length quartile is presented in Figure 7, showing that shorter comments exhibit a higher proportion of classification errors compared to longer comments. This finding indicates that transformer-based models operate more effectively when richer contextual information is available, as also reported in previous studies within the field of natural language processing.

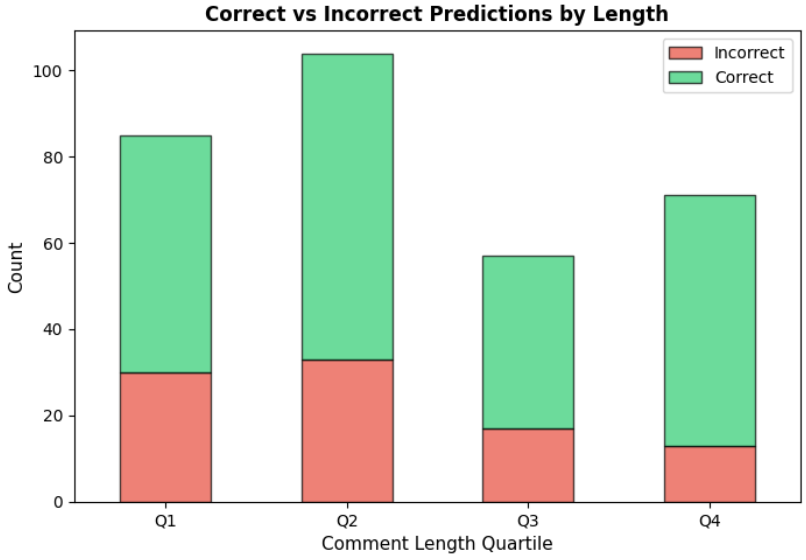

Figure 7. Distribution of Correct and Incorrect Predictions Based on Comment Length Quartiles

4. CONCLUSION

This study successfully developed an automated cyberbullying detection system for Indonesian-language TikTok comments using the IndoBERT model, achieving an accuracy of 70.66% and a ROC–AUC score of 0.7969. The model attained a recall of 0.7170 for the cyberbullying class, indicating adequate sensitivity in detecting harmful content an essential requirement in social media moderation, where failure to identify cyberbullying may lead to serious psychological consequences for victims. Error-based fairness analysis further indicates balanced performance across classes, with an accuracy gap of 2.08% and an Equal Opportunity Gap of 0.0208, both of which fall below the predefined thresholds. The relatively high Demographic Parity Gap (0.4132) is reported for completeness and should be interpreted cautiously due to the absence of demographic metadata and the inherently imbalanced linguistic characteristics of cyberbullying content. Additional analysis shows that classification performance improves as comment length increases, consistent with the contextual modeling capabilities of transformer architectures. Short comments with limited semantic context remain a major challenge, reinforcing the importance of contextual richness for reliable cyberbullying detection.

Despite these contributions, several limitations remain, including a relatively high false positive rate (30.38%), reduced performance on short comments, limited representation of regional dialects and evolving slang, and the exclusive focus on text-based signals. Future research should explore ensemble learning strategies, contextual augmentation for short texts, multimodal detection incorporating visual and emoji features, larger and more linguistically diverse datasets, pilot deployment studies in real moderation workflows, cross-platform transferability, and explainable AI techniques to strengthen human AI collaboration.

REFERENCES

[1] H. Dwistia, M. Sajdah, O. Awaliah, and N. Elfina, "Pemanfaatan Media Sosial Sebagai Media Pembelajaran Pendidikan Agama Islam," *Ar-Rusyd: Jurnal Pendidikan Agama Islam*, vol. 1, no. 2, pp. 81–99, 2022, doi: 10.61094/Arrusyd.2830-2281.33.

[2] D. McCashin and C. M. Murphy, "Using Tiktok for Public and Youth Mental Health – A Systematic Review and Content Analysis," *Clin. Child Psychol. Psychiatry*, vol. 28, no. 1, pp. 279–306, 2023, doi: 10.1177/13591045221106608.

[3] D. Keasaman, D. I. Pelabuhan Pengasinan, P. Jakarta, and Y. Mariah, "Jurnal Indonesia Sosial Sains," *Jurnal Indonesia Sosial Sains*, vol. 2, no. 3, p. 494, 2021.

[4] N. Rokhman, P. A. Maulan, and N. A. Wirahuda, "Analisis Penilaian Esai Secara Otomatis Menggunakan Natural Language Processing (NLP) dan Cosine Similarity," *Go Infotech: Jurnal Ilmiah Stmik Aub*, vol. 31, no. 1, pp. 41–52, 2025, doi: 10.36309/Goi.V31i1.359.

[5] T. Nugraha Manoppo and D. Hatta Fudholi, "Deteksi Cyberbullying Berdasarkan Unsur Perbuatan Pidana Yang Dilanggar Dengan Naive Bayes Dan Support Vector Machine," *Jurnal Sains Komputer & Informatika (J-Sakti)*, vol. 5, no. 1, pp. 10–19, 2021.

[6] F. Muftie, K. M. Yafi, and Q. M. Addina, "Perbandingan Performa Deteksi Cyberbullying Dengan Transformer, Deep Learning, Dan Machine Learning," *Jurnal Pendidikan Informatika Dan Sains*, vol. 13, no. 1, pp. 75–87, 2024, doi: 10.31571/Saintek.V13i1.4002.

[7] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian Multilabel Classification Using Indobert Embedding and Mbert Classification," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/Ijece.V14i1.Pp1071-1078.

[8] C. Denis, R. Elie, M. Hebiri, and F. Hu, "Fairness Guarantees in Multi-Class Classification with Demographic Parity," *Journal of Machine Learning Research*, vol. 25, pp. 1–46, 2024.

[9] A. Kurniasih and L. P. Manik, "On the Role of Text Preprocessing in Bert Embedding-Based Dnns for Classifying Informal Texts," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 927–934, 2022, doi: 10.14569/Ijacsa.2022.01306109.

[10] Hushian, "Cyberbullying Bahasa Indonesia, With Slang," *[Online]*. Available: https://www.kaggle.com/Datasets/Hushian/Cyberbullying-Dataset-With-Slang

[11] E. Küzeci, "Personal Data Protection Law," *Introduction to Turkish Business Law*, no. 016999, pp. 457–483, 2022.

[12] D. Rifaldi, Abdul Fadlil, and Herman, "Teknik Preprocessing pada Text Mining Menggunakan Data Tweet 'Mental Health,'" *Jurnal Pendidikan Teknologi Informasi*, vol. 3, no. 2, pp. 161–171, 2023.

[13] A. A. Khan, "Balanced Split: A New Train-Test Data Splitting Strategy for Imbalanced Datasets," *arXiv*, 2022.

[14] R. B. D. Figueiredo and H. A. Mendes, "Analyzing Information Leakage on Video Object Detection Datasets by Splitting Images into Clusters with High Spatiotemporal Correlation," *IEEE Access*, vol. 12, pp. 47646–47655, 2024, doi: 10.1109/Access.2024.3383047.

[15] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/Ijacsa.2024.0150235.

[16] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. Neurips, 2021.

[17] M. Riva, T. L. Parigi, F. Ungaro, and L. Massimino, "Hugging Face's Impact on Medical Applications of Artificial Intelligence," *Computational and Structural Biotechnology Reports*, vol. 1, no. March, p. 100003, 2024, doi: 10.1016/J.Csbr.2024.100003.

[18] Anugerah Simanjuntak et al., "Research and Analysis of Indobert Hyperparameter Tuning in Fake News Detection," *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, vol. 13, no. 1, pp. 60–67, 2024, doi: 10.22146/Jnteti.V13i1.8532.

[19] H. Tan, W. Shao, H. Wu, K. Yang, and L. Song, "A Sentence is Worth 128 Pseudo Tokens: A Semantic-Aware Contrastive Learning Framework for Sentence Embeddings," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, no. 2018, pp. 246–256, 2022, doi: 10.18653/V1/2022.Findings-Acl.22.

[20] L. Wu, G. Perin, and S. Picek, "I Choose You: Automated Hyperparameter Tuning for Deep Learning-Based Side-Channel Analysis," *IEEE Trans. Emerg. Top. Comput.*, vol. 12, no. 2, pp. 546–557, 2024, doi: 10.1109/Tetc.2022.3218372.

[21] S. Xie and Z. Li, "Implicit Bias of AdamW: ℓ∞-Norm Constrained Optimization," *Proc. Mach. Learn. Res.*, vol. 235, pp. 54488–54510, 2024.

[22] C. Wang, Y. Xiao, X. Gao, L. L. Li, and J. Wang, "Close the Gap Between Deep Learning and Mobile Intelligence by Incorporating Training in the Loop," *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, no. October 2019, pp. 1419–1427, 2019, doi: 10.1145/3343031.3350904.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[23] G. Alfonso-Francia et al., "Performance Evaluation of Different Object Detection Models for the Segmentation of Optical Cups and Discs," *Diagnostics*, vol. 12, no. 12, 2022, doi: 10.3390/Diagnostics12123031.

[24] D. Chicco and G. Jurman, "The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification," *Biodata Min.*, vol. 16, no. 1, Dec. 2023, doi: 10.1186/S13040-023-00322-4.

[25] K. C. Yuni K and I. Hanifuddin, "Analisis Fairness Terhadap Sistem Pembayaran Jasa Pengairan Sawah pada Petani Desa Bibrik Kecamatan Jiwan Kabupaten Madiun," *Journal of Economics, Law, and Humanities*, vol. 1, no. 2, pp. 59–74, 2022, doi: 10.21154/Jelhum.V1i2.1194.

[26] H. Al-Khalifa, K. Al-Khalefah, and H. Haroon, "Error Analysis of Pretrained Language Models (PLMs) in English-to-Arabic Machine Translation," *Human-Centric Intelligent Systems*, vol. 4, no. 2, pp. 206–219, 2024, doi: 10.1007/S44230-024-00061-7.