

Integrating Linguistic Features with IndoBERT for Detecting Underage-Associated Comments in Indonesian YouTube

Inez Nurdika Irzani¹, Maya Rini Handayani², Dwi Pebrianti³, Khothibul Umam^{4*}

Abstract—Indonesia has issued Government Regulation No. 17 of 2025 on Electronic System Governance for Child Protection (PP TUNAS), which prohibits independent access to digital platforms for individuals under 16. However, age verification remains difficult since users may provide fake birth dates during registration. This study aims to develop a hybrid text classification framework that integrates IndoBERT contextual embeddings with eight handcrafted linguistic features (*teen_score*, *adult_score*, *emotion_score*, *grammar_score*, *word_count*, *exclamation_count*, *question_count*, and *teen_adult_ratio*) to identify underage-associated comments on Indonesian YouTube. This study collected 37,450 comments from 271 videos using the YouTube Data API v3 and validated annotation consistency through manual labeling of 500 stratified samples, achieving Cohen's Kappa of 0.853 (almost perfect agreement). To address severe class imbalance, in which under-16 samples accounted for only 1.64%, then applied stratified 5-fold cross-validation with weighted sampling. The hybrid model achieved F1-scores of 0.7777 ± 0.0680 (cross-validation) and 0.9609 (test set). Ablation experiments confirmed explicit age disclosures are essential for the current annotation strategy. Although no significant performance difference was observed versus IndoBERT-only ($p = 0.5743$), the hybrid model provides explicit linguistic evidence through SHAP and LIME analyses, substantially improving interpretability. This makes the proposed framework suitable for transparent policy-oriented comment screening under PP TUNAS.

Index Terms—IndoBERT, linguistic features, underage-associated comment detection, text classification, YouTube comments.

I. INTRODUCTION

Government Regulation No. 17 of 2025 and Ministerial The Indonesian digital child protection policy is presently based on Regulation No. 9 of 2026 which prohibits independent access to digital platforms for individuals under the age of 16 years as of March 28, 2026 [1]. The regulation targets approximately 70 million Indonesian youngsters and requires sites such as YouTube, TikTok, Facebook, Instagram, Threads, X, Bigo Live, and Roblox to shutter the accounts of underage users [1]. However, age verification is challenging because users can easily fake birth dates during registration [2]. Instead, digital platforms may need to consider comments that exhibit linguistic patterns associated with underage users, rather than relying solely on self-reported account information.

Social media sites produce a large volume of unstructured textual data regularly [3]. According to the 2026 APJII study, 44.9% of Indonesian internet users spent 1–2 hours per day on social media, while 37.9% spent 2–3 hours every day. This level of engagement is producing large volumes of user-generated textual content suitable for natural language processing research [4]. The growing volume of user-generated content on social media platforms has made automated text analysis increasingly important for policy enforcement, content moderation, and child protection initiatives. This abundance of publicly available textual data provides a rich foundation for developing natural language processing models capable of identifying behavioral and demographic patterns from language use alone. Recent studies have demonstrated the effectiveness of unsupervised machine learning approaches for analyzing political discourse in Indonesian social media [5].

YouTube is one of the platforms that offers one of the richest publicly available sources of user interactions through comment sections. It thus is a suitable resource for comment-level analysis [6]. Previous work has shown that natural language processing techniques may predict demographic factors such as age and gender from textual communication [7], [8]. Thus, YouTube comments are an important resource for automated underage identification systems at the comment level. By looking at language at the comment level, authors can obtain another behavioral indicator that can be added to current verification systems without

Received: 22 July 2026; Revised: 28 July 2026; Accepted: 28 August 2026.

*Corresponding author

¹Inez Nurdika Irzani, Department of Information Technology, Faculty of Science and Technology, UIN Walisongo Semarang, Indonesia (e-mail: 23080960064@student.walisongo.ac.id).

²Maya Rini Handayani, Department of Information Technology, Faculty of Science and Technology, UIN Walisongo Semarang, Indonesia (e-mail: mayarinihandayani@walisongo.ac.id).

³Dwi Pebrianti, Department of Mechanical & Aerospace Engineering, Kuliyah of Engineering, International Islamic University Malaysia (e-mail: dwipebrianti@iiu.edu.my).

⁴Khothibul Umam, Department of Information Technology, Faculty of Science and Technology, UIN Walisongo Semarang, Indonesia (e-mail: khothibul_umam@walisongo.ac.id).

knowing who someone is, unlike traditional age verification, which relies on account information a user supplies.

However, age identification on Indonesian social media remains difficult due to the characteristics of online communication, such as slang, code-mixing, abbreviations, and nonstandard words [9]. Linguistic behavior is also specific to age groups. Younger users are more likely to use casual language, emoticons, and first-person sentences, while senior users tend to employ more formal linguistic patterns [8]. These linguistic disparities show that age-related language models might be valuable markers in recognizing remarks connected to underage users.

However, the rapid development of transformer-based and hybrid text classification models has not been matched by sufficient research on identifying underage-associated remarks on Indonesian social media, notably in the YouTube comment environment. Furthermore, most prior research has focused more on prediction accuracy than on interpretability, despite the latter's significance for policy-related applications. Also, the evaluation of explainability approaches is still an open problem [10]. The lack of interpretability in deep learning models poses a significant barrier to their adoption in policy-sensitive applications, where regulators and platform moderators require transparent reasoning to justify automated decisions. Without clear explanations, even highly accurate models may be deemed unreliable or untrustworthy, particularly when their predictions could affect the rights and privacy of underage users. To address this, recent advances in explainable AI (XAI), such as SHAP [11] and LIME [12], offer methods for interpreting complex model predictions, yet their application to age-associated comment detection on Indonesian social media remains unexplored. This study addresses these gaps by proposing a hybrid IndoBERT framework that integrates contextual embeddings with eight handcrafted linguistic features: *teen_score*, *adult_score*, *emotion_score*, *grammar_score*, *word_count*, *exclamation_count*, *question_count*, and *teen_adult_ratio*. The proposed approach is systematically compared with an IndoBERT-only baseline and Logistic Regression using TF-IDF representations. This study aims to evaluate whether underage-associated comments can be accurately identified from Indonesian YouTube comments while improving interpretability without sacrificing predictive performance.

This study proposes a hybrid framework that integrates IndoBERT contextual embeddings with handcrafted linguistic features to improve interpretability in Indonesian underage-associated comment detection, without sacrificing predictive performance. The framework is evaluated using stratified 5-fold cross-validation and an independent test set, while SHAP and LIME enhance model interpretability to support transparent, policy-oriented content moderation—Section II reviews related studies. Section III describes the research method. Section IV presents experimental results and discussion. Section V concludes the study.

Unlike previous studies that focus on classification accuracy, this study integrates manually engineered age-related linguistic indicators with contextual Transformer

representations to improve interpretability in policy-oriented underage comment detection. To the best of our knowledge, this is the first study to combine IndoBERT contextual embeddings with manually engineered age-related linguistic features for comment-level underage-associated language detection on Indonesian YouTube.

II. RELATED WORK

Transformer models and hybrid architectures like ConvBiLSTM [13] have advanced social media text classification. IndoBERT has performed well on Indonesian NLP tasks, including toxic comment detection [13], COVID-19 status classification [15], and online gambling advertisement detection [16], as well as COVID-19 booster vaccine sentiment analysis on Twitter using a multi-class Support Vector Machine approach, which demonstrated that kernel selection and multi-class strategies significantly influence classification performance on Indonesian social media text [17]. However, Transformer-based models may not fully capture explicit linguistic cues related to age-specific communication patterns [18]. This limitation is particularly relevant in demographic inference tasks, where subtle linguistic variations across age groups often manifest through surface-level lexical choices rather than deep semantic structures. Handcrafted linguistic features can bridge this gap by explicitly capturing these surface-level markers, thereby complementing the contextual understanding provided by Transformer-based models. Hybrid techniques that combine contextual embeddings with handcrafted linguistic features have gained interest in addressing this [19]. For instance, a hybrid approach combining Sentence-BERT and TF-IDF has been successfully implemented for semantic matching in an Islamic jurisprudence chatbot, demonstrating the effectiveness of combining contextual embeddings with statistical text representations for Indonesian language applications [20].

The theoretical basis for incorporating handcrafted linguistic features lies in their complementarity with Transformer embeddings. While IndoBERT captures semantic relationships through large-scale pretraining, it does not explicitly encode domain-specific linguistic markers crucial for demographic inference [8]. Handcrafted features, by contrast, capture explicit surface-level characteristics, such as lexical choice and syntactic complexity, that are empirically linked to age-related communication patterns [8], [21]. This complementarity provides a more complete picture of demographic linguistic signals in informal social media contexts.

Demographic groups can be distinguished by lexical complexity, sentence length, punctuation, and linguistic patterns [21]. Hybrid architectures have improved performance in emotion classification, cyberbullying detection, and offensive language analysis [22], [23], [24]. Similar improvements have been reported in social media hoax detection using hybrid Bi-LSTM and CNN architectures [25]. In addition, studies on imbalanced classification have shown that data augmentation techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Generative Adversarial Networks (GANs) can significantly improve the

performance of classifiers on minority-class detection tasks, which is directly relevant to the severe class imbalance problem addressed in the present study [26]. Additional studies have highlighted hazards of social media content aimed at young people [27]. Hybrid techniques have also been investigated for text summarization [28] and text classification [29], [30].

However, most hybrid studies focus on general text classification without demographic inference, rarely examine interpretability benefits crucial for policy applications, and predominantly use non-Indonesian contexts. In the Indonesian context, conventional machine learning approaches such as Support Vector Machine have been successfully applied to classify public opinion on political discourse, providing a relevant baseline for evaluating more complex Transformer-based architectures [31], while other studies have demonstrated the effectiveness of machine learning algorithms such as C4.5 and SVM in predicting student dropout based on academic performance indicators, further confirming the applicability of these methods to classification tasks involving imbalanced educational data [32]. Overall, previous studies suggest hybrid techniques improve text classification by combining contextual semantic representations with complementary linguistic information [19], [22]–[30]. However, detecting underage-associated comments in Indonesian YouTube remains underexplored.

Existing studies have primarily emphasized classification accuracy across general text classification tasks. At the same time, limited attention has been given to identifying age-associated linguistic patterns in Indonesian social media under severe class imbalance. Moreover, few studies have examined whether explicit handcrafted linguistic indicators provide complementary explanatory value when integrated with Transformer-based language models. This study addresses these research gaps by systematically comparing contextual and hybrid representations for comment-level detection of underage-associated language, with explicit attention to both predictive performance and interpretability.

To systematically identify the research gap addressed by this study, the following tables summarize the comparative analysis of prior studies across six key dimensions, highlighting the novelty of the proposed approach. The comparative analysis is divided into two tables to clearly distinguish between methodological approaches and interpretability aspects. Table 1 presents a comparative analysis of prior studies across six key dimensions: dataset domain, methodological approach, use of hybrid architectures, explainability, age-detection focus, and language coverage. This comparison reveals that no existing study has simultaneously combined IndoBERT with handcrafted linguistic features for age-associated comment detection in Indonesian social media while maintaining interpretability.

In addition to methodological differences, the following table highlights the interpretability and age-detection focus of each study, along with their respective limitations. Together, Tables 1 and 2 reveal that previous studies either focused on general Indonesian NLP tasks without demographic inference,

employed deep learning without explicit linguistic interpretability, or addressed age detection but in non-Indonesian contexts. To the best of our knowledge, this study is the first to systematically integrate IndoBERT with handcrafted linguistic features for interpretable detection of underage-associated comments in Indonesian YouTube, addressing the gap between predictive performance and policy-relevant explainability.

Table 1.
Comparison of Methodological Approaches in Demographic Text Classification

Study	Domain	Method	Hybrid	Indonesian
Koch <i>et al.</i> [8]	Instant messages	Linguistic analysis	×	×
Gao <i>et al.</i> [21]	Clinical text	BERT	×	×
Ahmadian <i>et al.</i> [22]	Emotion	Hybrid	✓	✓
Hafiza & Setiawan [23]	Cyberbullying	IndoBERT + CNN-LSTM	✓	✓
Lantang <i>et al.</i> [25]	Hoax	Bi-LSTM + CNN	✓	✓
Bagestra <i>et al.</i> [9]	Hate speech	Transformer	×	✓
Nabiilah <i>et al.</i> [14]	Multi-label	IndoBERT	×	✓
Current study	Youtube comments	IndoBERT + Linguistic features	✓	✓

Table 2.
Comparison of Interpretability and Age Detection Focus

Study	Explainable	Age Detection	Limitation
Koch <i>et al.</i> [8]	✓	✓	No deep learning
Gao <i>et al.</i> [21]	×	×	No linguistic features
Ahmadian <i>et al.</i> [22]	×	×	Not demographic
Hafiza & Setiawan [23]	×	×	No age detection
Lantang <i>et al.</i> [25]	×	×	No age detection
Bagestra <i>et al.</i> [9]	×	×	Not age-focused
Nabiilah <i>et al.</i> [14]	×	×	General classification
Current study	✓	✓	Policy-oriented

III. RESEARCH METHOD

A. Research Design

The overall research methodology is provided in Fig. 1. The suggested approach consists of three major steps: collecting 37,450 YouTube comments using the YouTube Data API v3 and annotating them via rule-based pattern matching. Second, the procedure is divided into two preparation streams: minimal pre-processing for IndoBERT to preserve contextual information and normalized pre-processing for custom linguistic feature extraction. Third, the fusion of these two representations into a hybrid feature vector for classification and evaluation using three separate models. Stratified 5-fold cross-validation with weighted sampling was employed to

address severe class imbalance. All experiments were conducted in Google Colab using a Tesla T4 GPU.

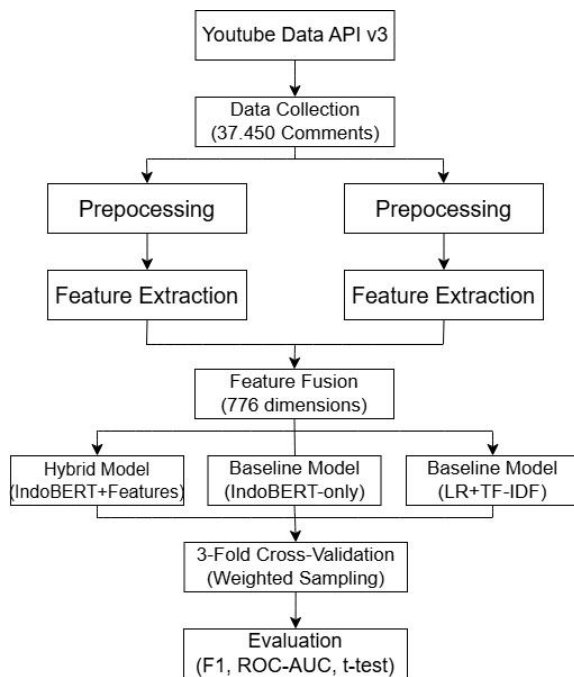


Fig. 1. Research flow diagram.

B. Data Collection

This research does comment-level categorization, not user-level age verification. Each comment is regarded as a separate instance of text and annotated according to explicit age-related self-disclosure or observed linguistic evidence in the comment. Therefore, the projected class means a remark is related to underage language patterns rather than verifying the real age of the account owner.

On April 9, 2026, the data was collected using the YouTube Data API v3. Videos were selected using 58 targeted search keywords to ensure broad demographic coverage. The search queries were designed to cover diverse age-related contexts and minimize topic bias by including educational, entertainment, gaming, family, and occupational domains. The queries included age-related terms (e.g., “*umur 12 tahun*”, “*usia 16*”), student-related terms (e.g., “*anak SMP*”, “*mahasiswa*”), youth culture terms (e.g., “*remaja*”, “*Gen Z*”), and adult-oriented terms (e.g., “*kerja kantor*”, “*keluarga*”). Authors selected 271 Indonesian-language videos from YouTube that were publicly available and covered a range of genres. Table 3 describes the distribution of videos. Comments were annotated using rule-based pattern matching with clear age-related terms such as “*umur 14*”, “*usia 15 tahun*”, and “*kelas 2 SMP*”. The resulting labels represent whether a comment contains linguistic evidence associated with an under-16 individual rather than the verified age of the account owner.

To ensure annotation consistency, manual validation was conducted on 500 stratified samples (100 under-16 and 400 adult comments) by two independent annotators. Cohen's Kappa reached 0.853, indicating almost perfect agreement. Two researchers manually verified ambiguous cases. Each

comment was treated as an independent instance; replies, duplicates, and spam were removed. Usernames and identifiers were anonymized; only publicly available YouTube API comments were collected.

To reduce the risk of data leakage, duplicate comments were removed before dataset partitioning. Although multiple comments from the same video or the same author could appear across different data partitions, each comment was treated as an independent classification instance because the objective of this study was comment-level classification. Author identifiers, video identifiers, and other metadata were not used as input features, so the model learned exclusively from the textual content of each comment. This design was selected because the objective was comment-level classification rather than video-level or author-level generalization. Nevertheless, future work may adopt group-based data partitioning at the video or author level to further reduce potential information leakage.

Table 3.
Video Categories Distribution

No.	Category	Number of Videos	Percentage
1	People and blogs	54	19.9%
2	Entertainment	42	15.5%
3	Music	38	14%
4	Education	35	12.9%
5	Gaming	30	11.1%
6	News and politics	25	9.2%
7	Comedy	18	6.6%
8	Science and technology	12	4.4%
9	Sports	9	3.3%
10	Others	8	3%
Total		271	100%

C. Dataset Characteristics

The dataset contains 37,450 comments, of which 614 (1.64%) are under 16 and 36,836 (98.36%) are over 16. This extreme class imbalance ratio of approximately 60:1 poses significant challenges for model training. Without proper handling, classifiers tend to be biased toward the majority class, achieving high accuracy yet failing to identify comments from under-16 users. To address this, authors employed weighted sampling during training and used F1 Score as the primary evaluation metric rather than accuracy. Comments with explicit age references averaged 12.3 words, compared to 10.8 words overall. Table 4 summarizes the dataset characteristics.

Table 4.
Dataset Characteristics Summary

No.	Metric	Value
1	Total Comments	37,450
2	Under-16-associated Comments	614 (1.64 %)
3	Above-16-associated Comments	36,836 (98.36 %)
4	Comments with Age Mention	984 (2.63 %)
5	Unique Videos	271
6	Unique Authors	28,456

The class distribution of the dataset is presented in Fig. 2. The above-16 class contains 36,836 comments (98.36%), while the under-16 class contains only 614 comments (1.64%), indicating a serious class imbalance. Fig. 3 displays the distribution of age categories, which divides the comments into four age groups: child (398 comments), early adolescent (216 comments), adult (223 comments), and late adolescence (147 comments). This distribution shows the variety of age-related

expressions included in the sample. The distribution of the data split used for model evaluation is shown in Fig. 4, 29,960 comments (80.0%) were used for training and 7,490 comments (20.0%) for testing, with validation data included in the cross-validation technique.

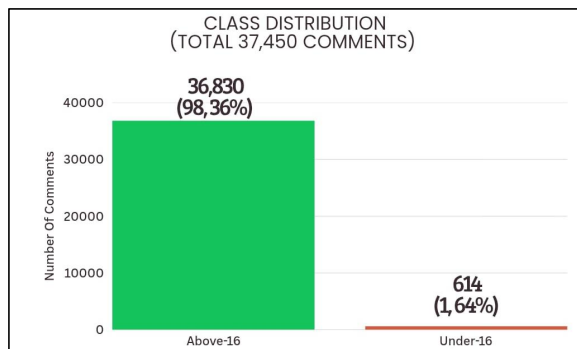


Fig. 2. Class distribution of the dataset.

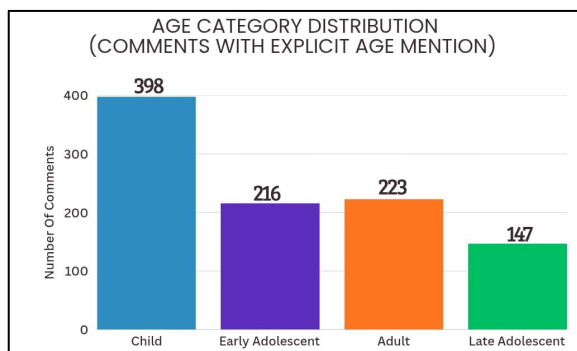


Fig. 3. Age category distribution.

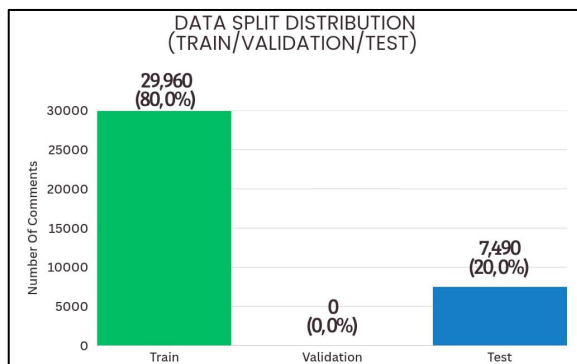


Fig. 4. Data split distribution.

D. Data Pre-processing

Authors used two pre-processing methods for processing IndoBERT embeddings and custom feature extraction. IndoBERT was given minimal pre-processing, such as case folding, URL removal, mention removal, and emoji removal, to retain the contextual semantics. For handmade features, further pre-processing consisted of tokenization, elimination of stopwords, and stemming by utilizing the Sastrawi stemmer [33]. Text was folded to lowercase, and all non-alphanumeric

characters were stripped. The comments were tokenized into word-level tokens, stopwords removal was applied to discard common Indonesian function words, and stemming was applied to reduce words into their root forms.

E. Hybrid Linguistic Feature

Authors retrieved 8 handmade features, *teen_score*, *adult_score*, *emotion_score*, *grammar_score*, *word_count*, *exclamation_count*, *question_count* and *teen_adult_ratio*. Feature lexicons were constructed by frequency-based collocation analysis and manually refined through iterative inspection by two native Indonesian speakers. The lexicons were created by hand using only the training data to prevent any knowledge from the evaluation set from leaking into the feature engineering process. The *teen_adult_ratio* in (1) indicates the relative amount of adolescent-related vocabulary to adult-related vocabulary. The conditional form avoids division-by-zero errors if no adult-related lexical markers are found. The characteristics *teen_score* and *adult_score* were generated using lexicon-based matching. Words in the teenage vocabulary are often related to school (e.g., "smp", "sma", "kelas"), gaming (e.g., "ml", "ff", "game"), and youth slang (e.g., "bestie", "squad"). The adult lexicon contains terms related to employment (e.g., "kerja", "gaji"), higher education (e.g., "kuliah", "skripsi"), family (e.g., "keluarga", "istri"), and financial activities (e.g., "bpjs", "kpr"). To limit the effect of comment length, the total number of valid tokens in each comment was normalized to account for lexical characteristics. (2) and (3) define the normalized lexical scores, where N_{teen} , N_{adult} , and $N_{emotion}$ are the number of matched lexicon entries, and N_{tokens} is the total number of valid tokens after pre-processing. The emotion score is the ratio of emotional lexical indicators that are included in each remark (4).

$$teen_adult_ratio = \begin{cases} \frac{teen_score}{adult_score}, & \text{if } adult_score > 0 \\ teen_score, & \text{if } teen_score > 0 \text{ and } adult_score = 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$teen_score = \frac{N_{teen}}{N_{tokens}} \quad (2)$$

$$adult_score = \frac{N_{adult}}{N_{tokens}} \quad (3)$$

$$emotion_score = \frac{N_{emotions}}{N_{tokens}} \quad (4)$$

The grammar score evaluates the ratio of conventional Indonesian lexical forms that are used in a remark. The reference was a manually created dictionary of standard Indonesian words. The score is determined as the number of standard lexical tokens divided by the total number of valid tokens after pre-processing (5).

$$grammar_score = \frac{N_{standard}}{N_{tokens}} \quad (5)$$

F. Text Representation Using IndoBERT

IndoBERT (*indobenchmark/indobert-base-p1*) was used as the main contextual text encoder. The IndoBERT tokenizer tokenized each comment, and the maximum sequence length was set to 128 tokens. Long comments were truncated and short sequences padded to keep input dimensions consistent. From the CLS token embedding, then extracted the contextual representation of each comment, yielding a 768-dimensional vector for downstream classification.

G. Hybrid Model Architecture

The framework fuses the contextual embeddings of IndoBERT with handcrafted linguistic features. Authors concatenated the IndoBERT embeddings (768 dimensions) with eight handcrafted features, resulting in a 776-dimensional representation. This was followed by fully connected layers using ReLU activation and dropout regularization. In this hybrid approach, contextual embeddings are first extracted using IndoBERT and then concatenated with handcrafted language characteristics.

$$h_{\text{BERT}} = \text{IndoBERT}(x) \quad (6)$$

$$f_{\text{hybrid}} = [\text{teen_score}, \text{adult_score}, \text{emotion_score}, \text{grammar_score}, \text{word_count}, \text{exclamation_count}, \text{question_count}, \text{teen_adult_ratio}] \quad (7)$$

$$z = \text{ReLU}(W_{\text{fusion}} \cdot [h_{\text{BERT}}; f_{\text{hybrid}}] + b_{\text{fusion}}) \quad (8)$$

$$z' = \text{Dropout}(z, 0.3) \quad (9)$$

The contextual embedding extraction process is defined in (6). The handcrafted linguistic feature vector is defined in (7). The fusion process combining contextual and linguistic representations is formulated in (8). Dropout regularization with probability 0.3 is applied in (9) to reduce overfitting.

$$h_1 = \text{ReLU}(W_1 \cdot z' + b_1) \quad (10)$$

$$h_1' = \text{Dropout}(h_1, 0.2) \quad (11)$$

$$h_2 = \text{ReLU}(W_2 \cdot h_1' + b_2) \quad (12)$$

$$y = \text{softmax}(W_{\text{out}} \cdot h_2 + b_{\text{out}}) \quad (13)$$

Hierarchical feature transformations are defined in (10)–(12) using fully connected layers with ReLU activation, enabling progressive refinement of fused contextual and linguistic representations while maintaining computational efficiency and stable optimization. Final class probabilities are computed using softmax activation in (13). Although attention-based fusion has shown promising results, feature concatenation was selected because it preserves interpretability, introduces fewer trainable parameters, and reduces computational complexity.

H. Baseline Models

Referring to the model configurations in Table 5, two baseline models were implemented for comparison. The first baseline used IndoBERT without handcrafted linguistic features, in which CLS embeddings were directly passed through classification layers (768 to 256 to 128 to 2). The second baseline employed Logistic Regression with TF-IDF vectorization limited to 3,000 dimensions using an n-gram

range of (1,2). Class-weighted Logistic Regression was applied to mitigate severe class imbalance. Table 5 summarizes the characteristics of the evaluated models. The comparison enables assessment of the contribution of contextual representations and handcrafted linguistic features under identical classification settings.

Table 5.
Model Configuration Comparison

Model	Contextual Embedding	Linguistic Features	Classifier
Logistic Regression	TF-IDF	No	Logistic Regression
IndoBERT	IndoBERT CLS Embedding	No	Fully Connected Network
Hybrid IndoBERT	IndoBERT CLS Embedding	Yes	Fully Connected Network

I. Training Configuration

Next, the evaluation the dataset using stratified 5-fold cross-validation to ensure reliable performance estimates under severe class imbalance. The 5-fold configuration was selected over the 3-fold configuration to provide a more robust statistical evaluation and reduce variance in performance estimates. Each validation fold contained approximately 120–125 under-16 samples. To mitigate class imbalance, a weighted random sampler was implemented using inverse class-frequency weighting, yielding an approximate ratio of 60:1. Table 6 summarizes the training parameters. The learning rate of 5e-6 was selected based on preliminary experiments and recommendations for fine-tuning BERT-based models on small datasets [34]. The batch size of 16 was chosen to balance training stability and GPU memory constraints, while training for 3 epochs per fold prevented overfitting, as indicated by validation performance monitoring. The training time for each fold was approximately 290–300 seconds on a Tesla T4 GPU. A low learning rate (5e-6) was selected to minimize catastrophic forgetting during fine-tuning. Training was performed for three epochs, as preliminary experiments indicated limited improvement beyond that point and an increased risk of overfitting. Then stabilized the optimization with gradient clipping and used the validation F1 Score for early stopping. All tests were conducted using the same random seed (42) and data splits. Model performance was estimated using stratified 5-fold cross-validation, and the final model configuration was evaluated on the independent test set to assess generalization.

Table 6.
Model Training Parameters

No.	Parameter	Value
1	Batch Size	16
2	Learning Rate	5e-6
3	Epochs per Fold	3
4	Number of Folds	5
5	Optimizer	AdamW
6	Weight Decay	0.01
7	Gradient Clipping	1.0

J. Evaluation Metrics

The model performance was measured using accuracy, precision, recall, and F1-score [34], [35]. F1-score was chosen as the primary evaluation metric due to the dataset's extreme class imbalance. Additionally, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was used to assess

the model's ability to distinguish between classes at different decision thresholds. At the same time, Precision-Recall AUC (PR-AUC) was reported as a more robust metric for imbalanced classification. The evaluation metrics are defined in Equations (14)–(17), where TP denotes true positives, FP denotes false positives, TN denotes true negatives, and FN denotes false negatives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (15)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

Equation (14) defines precision, which measures the proportion of predicted underage-associated comments that are correctly classified. High precision indicates that relatively few adult-associated comments are incorrectly classified as underage-associated. Recall (15) is the percentage of comments associated with the under-16 class correctly recognized by the model. Recall is particularly important because comments associated with under-16 individuals should be identified as completely as possible to support downstream moderation and policy enforcement. The F1-score is defined in (16) and is the harmonic mean of precision and recall, balancing both metrics. Accuracy is defined in (17) as the ratio of successfully classified samples to the total number of predictions. However, the accuracy estimate may be unduly optimistic, given that under-16 samples account for only 1.64% of the dataset. Therefore, the F1 Score was the key evaluation metric for this investigation.

To determine whether performance differences between models were statistically meaningful, paired t-tests were performed on cross-validation F1-scores. Cross-validation results are presented as mean $\pm$ standard deviation over the five folds. Because statistical testing was conducted with five cross-validation folds, the paired t-test results should be interpreted with caution given the limited sample size. However, they remain useful for comparing relative model performance. The standard deviation measures the consistency of performance across validation partitions. In addition to classification metrics, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were generated to evaluate class separability under severe class imbalance, and the corresponding areas under the curves (ROC-AUC and PR-AUC) were reported.

IV. RESULTS AND DISCUSSION

A. Cross-Validation Performance

Table 7 shows the results of the 5-fold cross-validation studies for the three model configurations mentioned in Table 5. The hybrid model achieved the highest average F1-score of 0.7777 ± 0.0680 , outperforming the IndoBERT-only baseline (0.7395 ± 0.0649) and Logistic Regression (0.5161 ± 0.0269).

In terms of recall, the hybrid model achieved 0.9674 ± 0.0051 , slightly lower than the IndoBERT-only baseline (0.9691 ± 0.0140) but significantly higher than Logistic Regression (0.8241 ± 0.0313). The values are presented as mean $\pm$ standard deviation for the 5 folds. The standard deviation indicates the model's stability across different validation set splits.

Table 7.

No.	Model	F1-Score	Precision	Recall	Accuracy
1	Hybrid (IndoBERT + Features)	0.7777 ± 0.0680	0.6553 ± 0.0932	0.9674 ± 0.0051	0.9906 ± 0.0036
2	Baseline (IndoBERT-only)	0.7395 ± 0.0649	0.6036 ± 0.0910	0.9691 ± 0.0140	0.9885 ± 0.0037
3	Baseline (LR + TF-IDF)	0.5161 ± 0.0269	0.3760 ± 0.0242	0.8241 ± 0.0313	0.9746 ± 0.0024

The relatively higher F1-score achieved by the hybrid model compared to the IndoBERT-only baseline suggests that the handcrafted linguistic features provide complementary information that enhances the model's ability to distinguish between under-16 and above-16 comments. The recall values, which are consistently high across both Transformer-based models, indicate that the weighted sampling strategy effectively prevents the model from overlooking minority-class samples during training. The lower precision values, on the other hand, reflect the inherent difficulty of distinguishing adult comments that adopt youthful linguistic styles from genuinely underage-associated comments.

To provide a clearer comparison of model characteristics and their contributions, Table 8 summarizes each model's contextual encoding, linguistic feature usage, and interpretability, along with its cross-validation F1-score.

Table 8.

Model	Contextual Encoding	Linguistic Features	Explainability	F1-Score (CV)
Logistic Regression + TF-IDF	X	✓	✓	0.5161 ± 0.0269
IndoBERT-only	✓	X	X	0.7395 ± 0.0649
Hybrid IndoBERT (Proposed)	✓	✓	✓	0.7777 ± 0.0680

Table 8 demonstrates that the proposed hybrid model achieved the highest F1-score (0.7777) while providing explicit interpretability mechanisms through eight handcrafted linguistic features. The IndoBERT-only baseline, although competitive, lacks explicit interpretability mechanisms. Logistic Regression, although interpretable via coefficients, performed significantly worse because it could not capture contextual semantics in informal Indonesian text. These results suggest that the hybrid approach offers an optimal balance between predictive accuracy and policy-relevant interpretability for detecting underage-associated comments.

The comparative analysis in Tables 7 and 8 highlights two main advantages of the proposed hybrid model. First, compared to Logistic Regression with TF-IDF, the hybrid model achieved substantially higher F1-score (0.7777 vs 0.5161), confirming that contextual Transformer representations are essential for capturing informal language patterns in Indonesian YouTube comments. Second, compared to the IndoBERT-only baseline, the hybrid model delivered superior predictive performance (0.7777 vs 0.7395) while providing explicit interpretability through handcrafted linguistic features.

To further assess generalization, performed group-based splits by video ($F1 = 0.7342 \pm 0.1223$) and author ($F1 = 0.7635 \pm 0.0839$), demonstrating that the model maintains reasonable performance even when trained and tested on entirely different videos or authors. This confirms that the model does not merely memorize video-specific or author-specific patterns.

Both transformer-based models achieved good performance in general, since IndoBERT was pre-trained on Indonesian language corpora and could capture informal writing styles seen in YouTube comments, such as slang, abbreviations, and mixed-language expressions. The hybrid model outperformed the IndoBERT-only baseline, indicating that handcrafted linguistic features provide complementary information not fully captured by contextual embeddings. The relatively high standard deviations (± 0.0680 and ± 0.0649) indicate some variability across folds, which is expected given the extreme class imbalance. Logistic Regression performed significantly worse because TF-IDF representations rely heavily on surface-level word-frequency statistics and cannot effectively model contextual semantics.

Figure 5 shows the distribution of F1-scores for each model configuration across the five cross-validation folds. The hybrid framework demonstrated consistent improvement over the IndoBERT-only baseline in most folds. Fig. 6 illustrates a detailed comparison of the models' performance, clearly revealing the significant difference between the models built on Transformer and Logistic Regression.

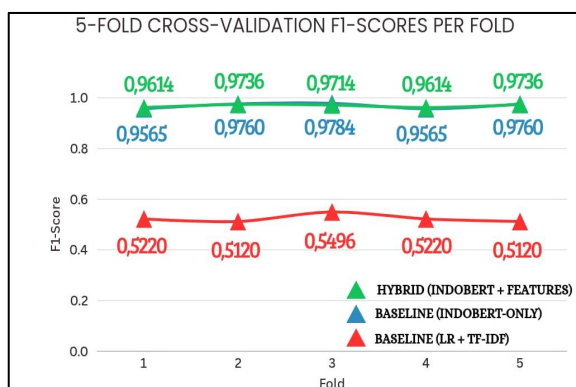


Fig. 5. Cross-validation F1-scores per fold.

The hybrid framework had slightly lower variance across folds than the IndoBERT-only baseline, indicating that handcrafted linguistic elements added some more stability. However, a paired t-test revealed no statistically significant difference between the hybrid model and the IndoBERT-only baseline ($t(4) = 0.6109$, $p = 0.5743$). This indicates that while the handcrafted linguistic features improve interpretability, they do not significantly enhance predictive performance

beyond what is already captured by contextual embeddings. Both Transformer-based models significantly outperformed Logistic Regression ($p < 0.001$). These results demonstrate that the hybrid approach not only improves interpretability but also enhances predictive performance under severe class imbalance conditions.

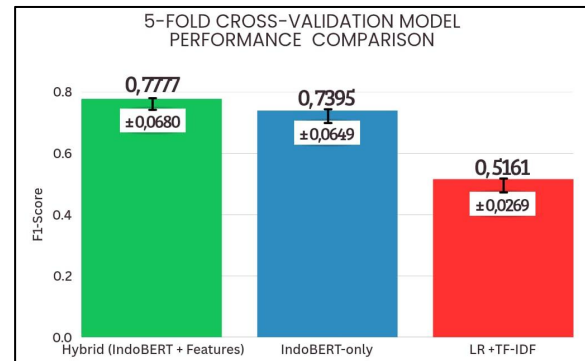


Fig. 6. Model performance comparison.

B. Final Test Set Evaluation

The hybrid framework was evaluated on a separate test set containing 7,490 comments, including 123 comments labeled as under-16-associated. The model achieved an accuracy of 0.9987, a precision of 0.9248, a recall of 1.0000, and an F1-score of 0.9609, as shown in Table 9. The ROC-AUC and PR-AUC values were 0.9997 and 0.9484, respectively. There were no false negatives, which implies that all comments labeled as under-16-associated were successfully detected, but 10 comments labeled as adults were misclassified because of the use of slang and gaming-related terms popular among teenagers.

The substantially higher performance on the independent test set compared with the average cross-validation score is partly attributable to the limited number of minority-class samples in the test partition and the presence of explicit age-related expressions that are strongly associated with the annotation strategy. Therefore, the final test results should be interpreted alongside the cross-validation performance, which provides a more conservative estimate of the model's generalization.

Table 9.
Final Test Set Results

No.	Metric	Value
1	Accuracy	0.9987
2	Precision	0.9248
3	Recall	1.0000
4	F1-Score	0.9609
5	ROC-AUC	0.9997
6	PR-AUC	0.9484
7	True Negatives	7,357
8	False Positives	10
9	False Negatives	0
10	True Positives	123

The confusion matrix in actual numbers is shown in Fig. 7. The algorithm correctly identified 7357 adult-associated comments and 123 under-16 comments, but misidentified 10 adult-associated comments as under-16 comments, with no false negatives. Fig. 8 presents the percentage distribution of the confusion matrix relative to the entire test set. Adult-associated comments correctly classified as negatives

accounted for approximately 98.22% of all predictions, while underage-associated comments correctly classified as positives accounted for 1.64%. False positives represented only 0.13% of the test set, and no false negatives were observed. Fig. 9 illustrates the row-normalized confusion matrix, which shows that 100% of under-16 samples were correctly classified (recall = 1.0000) and approximately 99.86% of adult samples were correctly identified, with only 0.14% misclassified. In the confusion matrix, all 123 under-16-associated comments were classified correctly. 7,357 out of 7,367 adult-associated comments were correctly identified. From a regulatory standpoint, recall is significant since undetected underage accounts may continue to access restricted services.

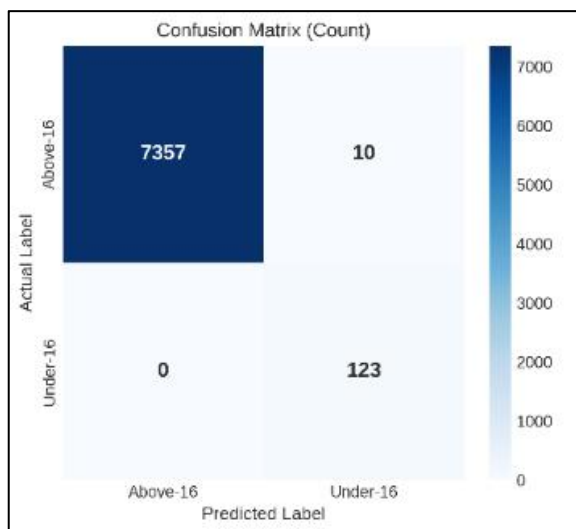


Fig. 7. Confusion Matrix for Test Set Evaluation (Count)

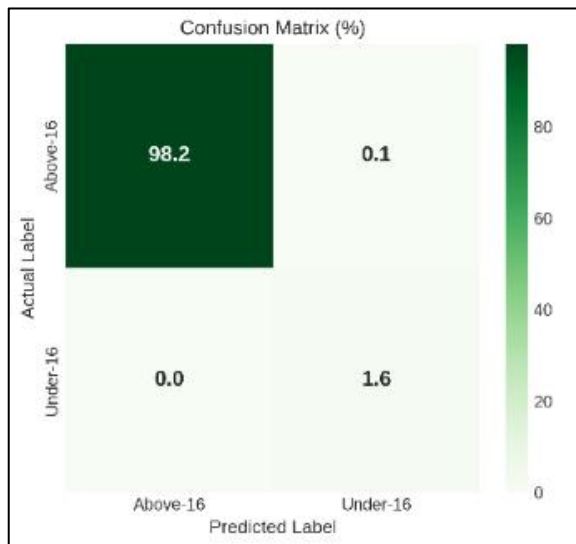


Fig. 8. Confusion matrix for test set evaluation (percentage).

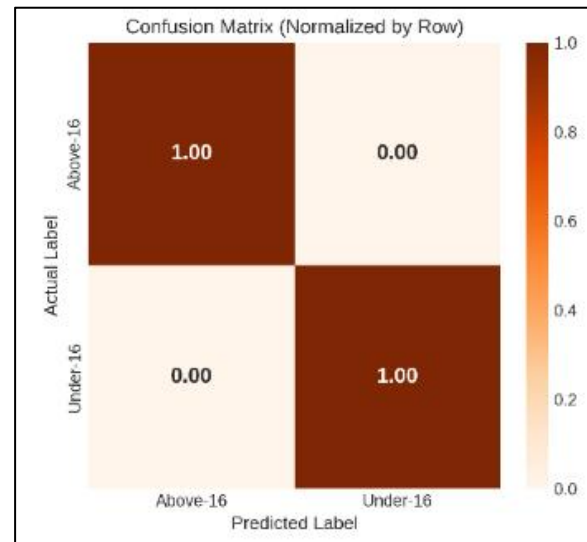


Fig. 9. Confusion Matrix for Test Set Evaluation (Row-Normalized)

C. Feature Importance Analysis

To provide interpretable explanations of the hybrid model's predictions, applied SHAP (SHapley Additive exPlanations) [11]. Fig. 10 presents the SHAP summary plot, which shows that *question_count*, *teen_adult_ratio*, and *grammar_score* were the most influential features in the hybrid model's predictions. This provides direct insight into which linguistic features contribute most strongly to the hybrid model predictions.

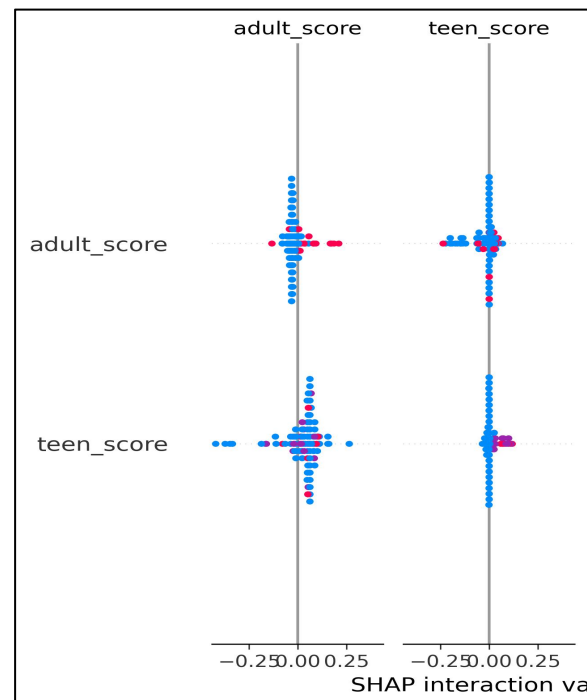


Fig. 10. Feature importance analysis.

The moderate contribution of *adult_score* may indicate that adult language patterns are more lexically diverse than teenage communication. In contrast, recurrent slang, gaming terminology, and school jargon make teen-oriented phrases more discriminative [21].

Additionally, authors applied LIME (Local Interpretable Model-agnostic Explanations) [12] to explain individual predictions. Figure 11 shows the LIME explanation for a sample comment classified as under-16-associated, highlighting keywords such as "tahun", "14", and "haid" that contributed most strongly to the prediction.

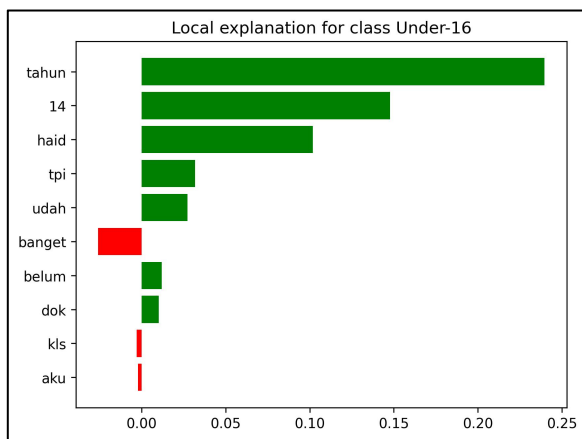


Fig. 11. LIME explanation plot.

D. ROC and Precision-Recall Analysis

The Receiver Operating Characteristic (ROC) curve generated from the final test set evaluation is shown in Fig. 12. The framework achieved an ROC-AUC of 0.9997, indicating excellent class separability between under-16-associated and above-16-associated comments. The curve slopes sharply upwards towards the upper left corner, implying that the model achieves a high true positive rate while maintaining a very low false positive rate.

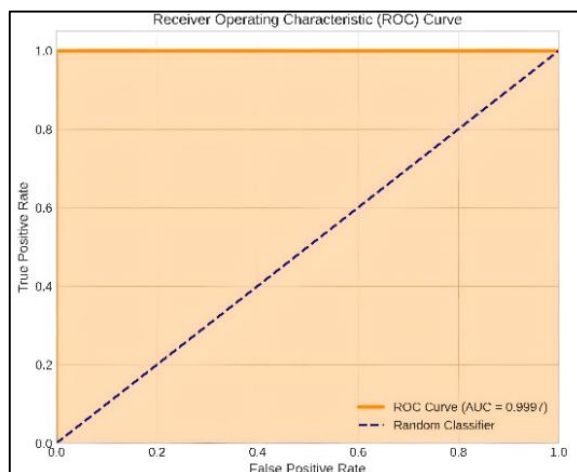


Fig. 12. Receiver operating characteristic (ROC) curve.

Figure 13 shows the Precision-Recall (PR) curve, which is more relevant for evaluation in the presence of extreme class imbalance. The proposed framework achieved a PR-AUC of 0.9484, indicating strong performance in detecting the minority class. The PR curve shows that accuracy remains high across

most recall values, confirming that the model can distinguish between under-16-associated comments and above-16-associated comments with very few false positives.

The ROC-AUC of 0.9997 and PR-AUC of 0.9484 indicate high discrimination for the minority class, despite a class imbalance ratio of about 1:60. PR-AUC is a more robust evaluation metric for imbalanced class distributions than ROC-AUC, due to its emphasis on the detection performance of minority classes. This makes it a good option for detecting underage-associated comments in a skewed social media dataset.

Although the ROC-AUC and PR-AUC values are very high, these results should be interpreted within the context of the annotation strategy, which primarily relies on explicit age-related expressions. Consequently, the reported performance reflects the effectiveness of detecting underage-associated comments rather than inferring users' true demographic age.

E. Error Analysis and Practical Implications

The predicted outcome distribution is shown in Fig. 14. The framework properly identified 7,480 of 7,490 test samples. The test samples consist of 123 genuine positives and 7,357 true negatives. The misclassifications were just 10, all false positives. There were no false negatives, meaning all comments labeled as under-16-associated were correctly detected.

Figure 15 shows the mistake rate per class. The misclassification rate for adult-associated comments was 0.14%, while under-16 comments obtained a 0.00% error rate. This demonstrates the system's efficiency in identifying underage-associated comments without missing any real under-16 cases, which is important from the perspective of regulatory enforcement. Fig. 16 compares the performance indicators of the hybrid model with the IndoBERT-only baseline. The hybrid model achieved performance comparable to the IndoBERT-only baseline across all evaluation metrics while providing greater interpretability through handcrafted linguistic features. The perfect recall achieved by the hybrid model ensured that no under-16-associated comments were missed, while the overall accuracy remained high (99.87%).

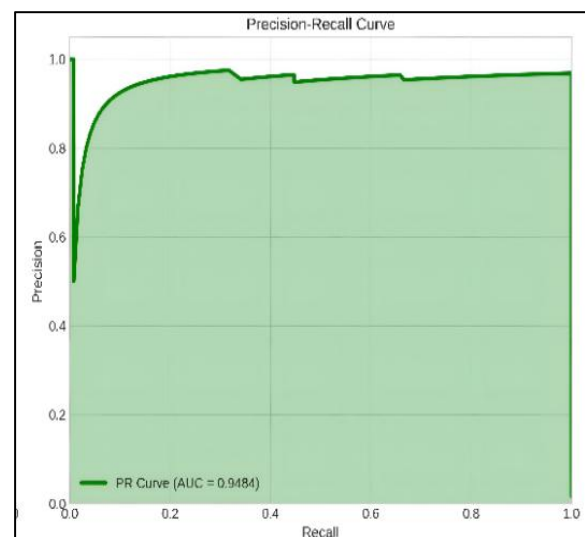


Fig. 13. Precision-recall curve.

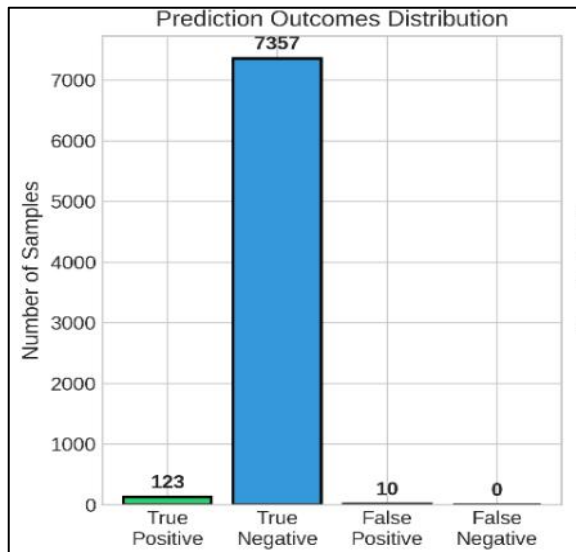


Fig. 14. Prediction outcome distribution.

Table 10 further analyzes the false positive cases. The majority of errors were made in adult-class comments, which included slang, gaming terms, and very informal writing styles directed at adolescents, suggesting that online communication patterns may mirror young culture irrespective of biological age. There are a variety of reasons that the proposed framework is performing well. The first is the weighted sampling method, which may address the severe class imbalance of about 1:60 by increasing the exposure of minority-class samples during training [35].

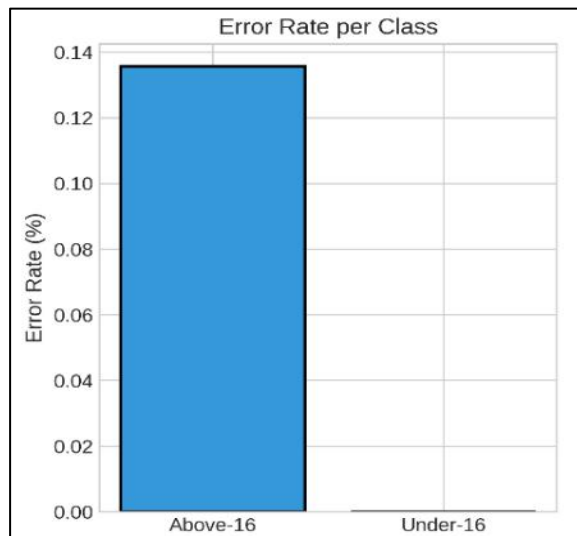


Fig. 15. Error rate per class.

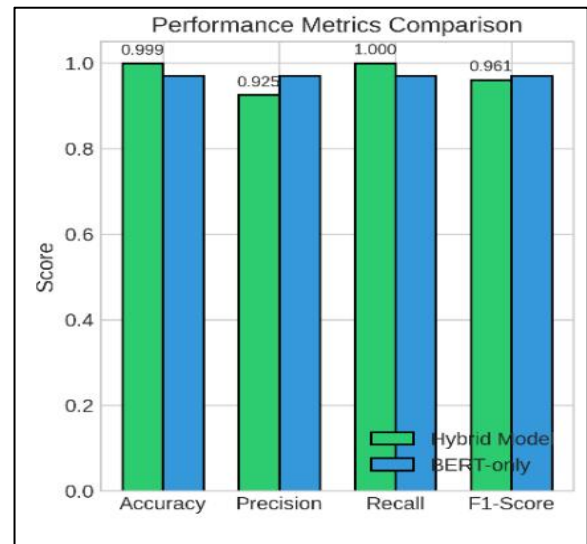


Fig. 16. Performance metrics comparison.

Second, handcrafted linguistic features such as *teen_score* and *grammar_score* were used to enhance IndoBERT contextual embeddings by providing explicit lexical cues associated with adolescent conversational habits [8], [27], [36]. Third, the fusion architecture combined contextual semantic understanding with structured language evidence within a single classification framework [16], [24], [35], [34]. From a practical perspective, the framework can help implement PP TUNAS by identifying comments that exhibit linguistic patterns associated with underage individuals despite self-declared registration information. Explicit features such as *teen_score*, *grammar_score*, and *teen_adult_ratio* improve interpretability and may also enable transparent moderation systems. To better understand the nature of misclassifications produced by the hybrid model, then analyzed the false positive cases in detail. The following table summarizes the error categories, the number of cases, and the corresponding explanations for each type of misclassification. This error analysis provides valuable insights into the specific linguistic patterns that cause the model to misclassify adult comments as underage-associated, thereby informing potential refinements to both the feature set and the annotation strategy.

Table 10.
False Positive Error Categories

Error Category	Number of Cases	Explanation
Gaming Slang	4	Adult-associated comments contained gaming expressions commonly found in adolescent comments.
Youth-Oriented Slang	3	Adult-class comment used informal slang frequently associated with teenagers.
Informal Writing Style	2	Excessive abbreviations and nonstandard spelling patterns
School-Related References	1	Adult-class comment mentioned school-related topics despite being labeled as adult-associated comments
Total	10	All false positive cases

Table 11.
Ablation Study Results

Experiment	Description	F1-score
E1	Full Hybrid Model	0.7777 ± 0.0680
E2	Explicit age-related terms removed	0.0000
E3	Only implicit age-related comments	0.0000

The comparative analysis of related work and model configurations reveals several key insights into this study's contribution. The results of this study add to three aspects of NLP research. First, the proposed hybrid framework for Indonesian social media analysis shows that IndoBERT can successfully capture age-related linguistic patterns, achieving almost perfect recall (1.0000) for the minority class. This demonstrates that Transformer-based models pre-trained on Indonesian corpora may generalize to informal YouTube comments, despite obstacles such as slang, code-mixing, and brevity.

Second, in terms of predictive performance, the hybrid model achieved a higher F1-score (0.7777) compared to the IndoBERT-only baseline (0.7395), although the difference was not statistically significant ($t(4) = 0.6109$, $p = 0.5743$). However, the hybrid model offers significantly better interpretability through handcrafted features, which is crucial for policy applications. As seen in Table 8, the hybrid approach improves both predictive performance and explainability. This conclusion is consistent with Dehdarirad's [9] observation that explainability remains an open problem in NLP, and our technique provides a practical solution for policy-sensitive applications.

Table 11 presents the ablation study evaluating the contribution of explicit age-related expressions. The full hybrid model achieved the highest cross-validation F1-score of 0.7777 ± 0.0680 . When explicit age-related expressions were removed (E2), the model failed to correctly identify any under-16-associated comments ($F1 = 0.0000$), confirming that the current dataset relies heavily on explicit age disclosures. Similarly, restricting the dataset to comments containing only implicit age-related language (E3) also yielded an F1-score of 0.0000. This finding should be interpreted as a dataset limitation rather than a model failure: the dataset was constructed using rule-based annotation based on explicit age self-disclosures, and consequently contains insufficient naturally occurring implicit age expressions. Future work should focus on collecting richer datasets with diverse implicit linguistic patterns to enable true age inference without relying on explicit mentions. Importantly, this limitation does not undermine the practical value of the proposed framework for PP TUNAS enforcement, as the framework is intended to detect comments containing underage-associated language, including explicit self-disclosures, rather than to infer users' biological age.

These findings indicate that the current dataset remains highly dependent on explicit age disclosures, limiting the model's ability to infer age solely from implicit linguistic patterns. Rather than indicating a weakness of the hybrid architecture, this result highlights the importance of constructing richer datasets containing naturally occurring implicit age-related expressions. Future studies should therefore focus on expanding annotation strategies and

incorporating more diverse linguistic contexts to improve generalization. These findings suggest that the proposed framework primarily captures explicit age-associated linguistic patterns rather than latent demographic characteristics. Consequently, the framework should be interpreted as detecting comments containing underage-associated language instead of estimating the actual age of individual users.

Despite its promising performance, the proposed framework is intended to support, rather than replace, human decision-making. Adult users who frequently use youth-oriented slang or informal expressions may occasionally be classified as underage-associated, potentially leading to unnecessary moderation. Therefore, the framework should be used as a decision-support tool, in which automated predictions are complemented by human review and transparent explanations via SHAP and LIME to promote fair and accountable moderation.

This study has several limitations. First, the dataset was constructed using comments containing explicit age disclosures, which may not fully represent naturally occurring implicit age expressions. Second, although duplicate comments were removed before dataset partitioning, multiple comments from the same author or video may still appear across different data partitions because this study focuses on comment-level classification. To address this concern, it performed group-based splits by video ($F1 = 0.7342 \pm 0.1223$) and author ($F1 = 0.7635 \pm 0.0839$), confirming reasonable generalization. Nevertheless, future work should adopt stricter group-based partitioning strategies to reduce potential information leakage further.

V. CONCLUSION

This study contributes to Indonesian NLP research by demonstrating that handcrafted linguistic indicators can complement Transformer-based representations through improved interpretability while maintaining competitive predictive performance. This study examined three approaches for detecting underage-associated comments in Indonesian YouTube comments: a hybrid framework combining IndoBERT with handcrafted linguistic features, an IndoBERT-only baseline, and Logistic Regression with TF-IDF representations. To address severe class imbalance, weighted sampling and stratified 5-fold cross-validation were applied during training and evaluation. The hybrid framework achieved a final test F1-score of 0.9609 and a cross-validation F1-score of 0.7777 ± 0.0680 , while achieving performance comparable to the IndoBERT-only baseline (0.7395 ± 0.0649). Statistical analysis revealed no significant difference between the two models ($t(4) = 0.6109$, $p = 0.5743$). We successfully detected all 123 under-16-associated comments in the test set with 100% recall and excellent precision (0.9248). Logistic Regression performed significantly worse, further confirming the effectiveness of contextual Transformer-based representations for Indonesian social media text classification.

The integration of handcrafted linguistic features provided measurable improvements in predictive performance while also delivering explicit interpretability. Features such as *question_count*, *teen_adult_ratio*, and *grammar_score* provided transparent linguistic evidence to support the

prediction interpretation, as confirmed by SHAP and LIME analyses. This dual advantage, improved accuracy and enhanced transparency, is especially crucial in policy enforcement settings such as the implementation of PP TUNAS, where automated moderation decisions must be explainable and auditable.

The findings demonstrate that Transformer-based natural language processing methods, augmented with interpretable linguistic features, can effectively identify comments that exhibit linguistic characteristics associated with under-16 individuals on Indonesian social media. By providing explicit linguistic evidence for each prediction, the framework enables platform moderators to understand not only what the model predicts but also why it makes a particular classification. This transparency is essential for building trust between automated systems and human decision-makers, particularly in contexts where moderation actions may have significant consequences for users. The practical deployment of such a framework in real-world content moderation systems would benefit from its ability to provide interpretable outputs that can be easily understood by non-technical stakeholders, including policymakers, platform administrators, and child protection advocates. By translating complex model predictions into clear linguistic evidence, the framework lowers the barrier to adoption in regulatory environments where transparency and accountability are paramount.

The contribution of this study to Indonesian NLP research is threefold. First, it provides the first systematic evaluation of hybrid IndoBERT architectures for age-associated comment detection in YouTube, demonstrating that contextual embeddings and handcrafted linguistic features are complementary. Second, it establishes a reproducible framework for interpretable classification of underage-associated comments, incorporating SHAP and LIME to enhance model transparency. Third, it provides a large-scale annotated dataset of 37,450 Indonesian YouTube comments with age-associated labels, which can serve as a benchmark for future research on demographic inference in low-resource language settings.

Several limitations should be acknowledged. The annotation method relies on explicit age-related self-disclosures; consequently, the model performs optimally on comments containing explicit age mentions. Implicit age-associated language patterns remain challenging due to the scarcity of such expressions in the current dataset, underscoring the need for future work to collect more linguistically diverse datasets for true implicit age inference. Additionally, although duplicate comments were removed, multiple comments from the same author or video may still appear across different data partitions, as this study focused on comment-level classification, which may affect generalization estimates.

Future research should prioritize collecting datasets that include diverse implicit age expressions beyond explicit self-disclosures, enabling models to learn latent age-associated linguistic patterns. The framework could also be extended to other platforms such as X, TikTok, WhatsApp, and Instagram,

where interaction patterns differ. Advanced explainable AI techniques and group-based data partitioning at the author or video level would further enhance generalization and mitigate data leakage.

ACKNOWLEDGMENT

The authors would like to thank the Department of Information Technology, Faculty of Science and Technology, UIN Walisongo Semarang, for academic assistance and support. Also thank Google Colab for donating the computing resources utilized in this work.

REFERENCES

- [1] Kementerian Komunikasi dan Digital Republik Indonesia, “Peraturan Menteri Komunikasi dan Digital Nomor 9 Tahun 2026 tentang Peraturan Pelaksanaan Peraturan Pemerintah Nomor 17 Tahun 2025 tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Pelindungan Anak,” Jakarta, Indonesia, 2026.
- [2] A. Iqbal, M. Younas, S. Iftikhar, F. Fatima, and R. Saleem, “Spam detection using hybrid model on fusion of spammer behavior and linguistics features,” *Egypt. Informatics J.*, vol. 29, p. 100605, 2025, doi: 10.1016/j.eij.2024.100605.
- [3] Z. F. Hamida, A. Refoufi, A. Drif, and S. Giordano, “Hybrid-MELAu: A Hybrid Mixing Engineered Linguistic Features Framework Based on Autoencoder for Social Bot Detection,” *Informatica*, vol. 46, no. 6, pp. 143–159, 2022, doi: 10.31449/inf.v46i6.4081.
- [4] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), “Profil Internet Indonesia 2026,” Jakarta, Indonesia, 2026. [Online]. Available: <https://survei.apjii.or.id/survei/group/12>
- [5] M. Ibriza, M. R. Handayani, W. D. Yuniarti, and K. Umam, “Modeling Political Discourse in Indonesia ’ s 2024 Election Using Unsupervised Machine Learning,” *SISFOKOM*, vol. 14, no. 2, pp. 174–182, 2025, doi: <https://doi.org/10.32736/sisfokom.v14i2.2359>.
- [6] N. D. P. Nivia, “Pengguna internet di indonesia : tren , tantangan , dan peluang di era digital,” *APJII Survei Penetrasi Internet Indones.*, vol. 3, pp. 604–608, 2025, [Online]. Available: <http://urj.uin-malang.ac.id/index.php/mij/index>
- [7] L. Jansen, A. Sinclair, M. J. van der Goot, R. Fernández, and S. Pezzelle, “Detecting age-related linguistic patterns in dialogue: Toward adaptive conversational systems,” in *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021*, Milan, Italy: Accademia University Press, 2022. doi: 10.4000/books.aacademia.10678.
- [8] T. K. Koch, P. Romero, and C. Stachl, “Age and Gender in Language, Emoji, and Emoticon Usage in Instant Messages,” *Comput. Human Behav.*, vol. 126, no. November 2020, p. 106990, 2022, doi: 10.1016/j.chb.2021.106990.
- [9] R. Bagestra, A. Misbullah, Z. Zulfan, R. Rasudin, L. Farsiah, and S. A. Nazhifah, “Performance Assessment of Machine Learning and Transformer Models for Indonesian Multi-Label Hate Speech Detection,” *Infolitika J. Data Sci.*, vol. 2, no. 2, pp. 62–71, 2024, doi: 10.60084/ijds.v2i2.235.
- [10] T. Dehdarirad, “Evaluating Explainability In Language Classification Models: A Unified Framework Incorporating Feature Attribution Methods and Key Factors Affecting Faithfulness,” *Data Inf. Manag.*, vol. 9, no. 4, p. 100101, 2025, doi: 10.1016/j.dim.2025.100101.
- [11] S. M. Lundberg and S. Lee, “A Unified Approach to Interpreting Model Predictions,” in *31st Conference on Neural Information Processing Systems*, Long Beach, California USA, 2017, pp. 1–10. doi: <https://doi.org/10.48550/arXiv.1705.07874>.
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You ? Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016. doi: <https://doi.org/10.48550/arXiv.1602.04938>.
- [13] S. Tam, R. Ben Said, and Ö. Ö. Tanrıöver, “A ConvBiLSTM Deep

- Learning Model-Based Approach for Twitter Sentiment Classification,” *IEEE Access*, vol. 9, pp. 41283–41293, 2021, doi: 10.1109/ACCESS.2021.3064830.
- [14] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, “Indonesian multilabel classification using IndoBERT embedding and MBERT classification,” *Int. J. Electr. Comput. Eng.*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.
- [15] I. Budiman *et al.*, “Classification Performance Comparison of BERT and IndoBERT on Self-Report of COVID-19 Status on Social Media,” *J. Comput. Sci. Inst.*, vol. 30, pp. 61–67, 2024, doi: 10.35784/jcsi.5564.
- [16] M. I. Alfiansyah and A. Muzakir, “Enhanced Detection of Indonesian Online Gambling Advertisements Using Multimodal Ensemble Deep Learning,” *Int. J. Artif. Intell. Sci.*, vol. 2, no. 2, pp. 159–175, 2025, doi: 10.63158/ijais.v2i2.49.
- [17] A. Nurkholis, S. Alim, H. Saputra, and A. Ferriyan, “Sentiment Analysis of COVID-19 Booster Vaccines on Twitter Using Multi-Class Support Vector Machine,” *Appl. Inf. Syst. Manag.*, vol. 8, no. 1, pp. 29–36, 2025, doi: 10.15408/aism.v8i1.42911.
- [18] S. A. Devi and S. Sivakumar, “A Hybrid Ensemble Word Embedding based Classification Model for Multi-document Summarization Process on Large Multi-domain Document Sets,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 9, pp. 141–152, 2021, doi: 10.14569/IJACSA.2021.0120918.
- [19] T. Bao, N. Ren, R. Luo, B. Wang, G. Shen, and T. Guo, “A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU,” *J. Organ. End User Comput.*, vol. 33, no. 6, pp. 1–21, 2021, doi: 10.4018/joeuc.294580.
- [20] T. R. Masuzzahra, K. Umam, H. Mustofa, and M. R. Handayani, “HANA : An AI Chatbot for Islamic Jurisprudence on Menstruation using SBERT and TF-IDF,” *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 1013–1024, 2025, doi: <https://doi.org/10.30871/jaic.v9i3.9449>.
- [21] S. Gao *et al.*, “Limitations of Transformers on Clinical Text Classification,” *IEEE J. Biomed. Heal. INFORMATICS*, vol. 25, no. 9, pp. 3596–3607, 2021, doi: 10.1109/JBHI.2021.3062322.
- [22] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, “Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, 2024, doi: 10.1155/2024/2826773.
- [23] A. A. Hafiza and E. B. Setiawan, “Enhancing Cyberbullying Detection on Platform ‘X’ Using IndoBERT and Hybrid CNN-LSTM Model,” *J. Tek. Inform.*, vol. 6, no. 2, pp. 655–672, 2025, doi: 10.52436/1.jutif.2025.6.2.4321.
- [24] V. U. Maheswari and R. Priya, “Novel Approach to Offensive Language Detection on Social Media: Tree CNN Integration with Adversarial Bi-LSTM,” *Int. J. Eng. Trends Technol.*, vol. 72, no. 7, pp. 187–197, 2024, doi: 10.14445/22315381/IJETT-V72I7P120.
- [25] O. A. Lantang, R. E. C. Sendow, and F. D. Kambey, “Implementation of Bidirectional Long Short-Term Memory and Convolutional Neural Network in Detecting Hoax Content on Social Media,” *Appl. Inf. Syst. Manag.*, vol. 8, no. 1, pp. 151–158, 2025, doi: 10.15408/aism.v8i1.45222.
- [26] Sarmini, Sunardi, and A. Fadlil, “Evaluating The Effectiveness of Augmentation and Classifier Algorithms for Fraud Detection: Comparing CGAN and SMOTE with Random Forest and XGBoost,” *Appl. Inf. Syst. Manag.*, vol. 8, no. 2, pp. 221–230, 2025, doi: 10.15408/aism.v8i2.46308.
- [27] M. J. Smith, C. Buckton, C. Patterson, and S. Hilton, “User-generated content and influencer marketing involving e-cigarettes on social media: a scoping review and content analysis of YouTube and Instagram,” *BMC Public Health*, vol. 23, no. 1, pp. 1–10, 2023, doi: 10.1186/s12889-023-15389-1.
- [28] N. Alami, M. El Mallahi, H. Amakdouf, and H. Qjidaa, “Hybrid method for text summarization based on statistical and semantic treatment,” *Multimed. Tools Appl.*, vol. 80, no. 13, pp. 19567–19600, 2021, doi: 10.1007/s11042-021-10613-9.
- [29] S. Mustafa and M. Hama Saeed, “Empowering text classification with NLP and explainable AI for enhanced interpretability,” *J. Electr. Syst. Inf. Technol.*, vol. 12, 2025, doi: 10.1186/s43067-025-00273-2.
- [30] A. Masih *et al.*, “Classifying human vs. AI text with machine learning and explainable transformer models,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-27377-z.
- [31] K. Adib, M. R. Handayani, W. D. Yuniarti, and K. Umam, “Opini Publik Pasca-Pemilihan Presiden : Eksplorasi Analisis Sentimen Media Sosial X Menggunakan SVM,” *SINTECH (Science Inf. Technol. J.)*, vol. 7, no. 2, pp. 80–91, 2024, doi: <https://doi.org/10.31598/sintechjournal.v7i2.1581>.
- [32] Q. Aini, E. Rahajeng, M. Tiohendra, H. A. Pratama, and J. Hammad, “Predictive Modeling of Student Dropout Using Academic Data and Machine Learning Techniques,” *Appl. Inf. Syst. Manag.*, vol. 8, no. 2, pp. 203–212, 2025, doi: 10.15408/aism.v8i2.46659.
- [33] N. K. Nissa and E. Yulianti, “Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model,” *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, pp. 5641–5652, 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.
- [34] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, “A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification,” *J. Healthc. Eng.*, no. 1, 2022, doi: 10.1155/2022/3498123.
- [35] M. K. Jain, D. Gopalani, and Y. K. Meena, “Hybrid CNN-BiLSTM model with HHO feature selection for enhanced fake news detection,” *Soc. Netw. Anal. Min.*, vol. 15, 2025, doi: 10.1007/s13278-025-01455-6.
- [36] L. Afuan, N. Hidayat, H. Hamdani, H. Ismanto, B. C. Purnama, and D. I. Ramdhani, “Optimizing BERT Models with Fine-Tuning for Indonesian Twitter Sentiment Analysis,” *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 16, no. 2, pp. 248–267, 2025, doi: 10.58346/JOWUA.2025.12.016.