Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

Comparative Analysis of IndoBERT and mBERT for Online Gambling Comment Detection in Indonesian Social Media

Satria Adi Nugraha[1], Citra Lestari[*2], Karyna Budi Sanjaya[3], Rafi Abhista Naya[4], Jocelyn Jolie[5]

[1,2,3,4,5]School of Information Technology, Universitas Ciputra, Indonesia

Email: [2]caecilia.citra@ciputra.ac.id

Received : Jan 29, 2026; Revised : Feb 23, 2026; Accepted : Mar 26, 2026; Published : Apr 20, 2026

Abstract

The rapid growth of illegal online gambling promotions in Indonesian social media comments requires automated detection systems capable of handling informal and noisy text. This study aims to evaluate the effectiveness of Transformer-based language models for detecting online gambling-related comments in Indonesian Twitter and YouTube data. Two pre-trained models, IndoBERT and mBERT, were fine-tuned and compared using a labeled dataset consisting of gambling and non-gambling comments. Model performance was evaluated using accuracy, precision, recall, and F1-score. Experimental results show that IndoBERT achieved 98% accuracy and F1-score, outperforming mBERT, which achieved 96% on the same dataset. Additionally, performance was compared against a recurrent neural network baseline to validate the effectiveness of Transformer-based architectures. The findings demonstrate that language-specific pre-training provides measurable advantages for detecting domain-specific content in Indonesian social media. This study contributes empirical evidence supporting the application of Transformer models for automated moderation of harmful online content in Indonesian digital platforms.

Keywords : *BERT, Cybercrime Detection, IndoBERT, MBERT, Social Media Analysis, Transformer Models.*

1. INTRODUCTION

In recent years, online gambling has emerged as a growing social and economic concern in Indonesia. According to a report by Liputan6.com, approximately 8.8 million Indonesians were engaged in online gambling activities in 2024 [1]. The widespread accessibility of online platforms has contributed to the rise of gambling addiction, which poses serious economic risks. Syakira et al. [2] evealed that individuals involved in online gambling tend to deplete their financial resources on gambling-related expenditures. Similarly, Laras et al., as cited by Syakira et al. [2], observed that gambling addiction often reduces productivity as individuals devote excessive time and energy to gambling activities. Such behaviors frequently lead to severe financial losses and jeopardize the long-term financial stability of affected households [3][4][5].

Beyond its economic implications, online gambling also produces profound social and psychological consequences. A qualitative study reported that individuals with direct experience in online gambling are prone to social withdrawal and increased family conflicts [3]. Moreover, symptoms of anxiety, depression, and stress tend to intensify among habitual gamblers. A recent investigation focusing on university students, both with and without disabilities, demonstrated that gambling involvement correlates with significant psychological distress, interference in daily routines, and heightened vulnerability among individuals with disabilities [6].

The increasing prevalence of online gambling is closely tied to the growing exposure to gambling-related advertisements across digital platforms. A national survey by Populix in 2023 involving 1,058 respondents found that 82% of Indonesians had encountered online gambling advertisements, with 63%

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

reporting exposure nearly every time they accessed the internet [7]. The majority of these advertisements appeared on popular social media platforms such as Instagram (46%), YouTube (45%), and Facebook (45%) [7]. The persistence of such advertisements is largely due to their deceptive presentation, often disguised as legitimate gaming promotions or influencer endorsements. Studies have shown that gambling advertisers increasingly employ subtle promotional strategies, including undisclosed sponsorships and the use of influencer-driven narratives, to normalize gambling-related content [8][9][10]. Since late 2024, gambling advertisements have begun infiltrating comment sections on YouTube, both in live-streamed and pre-recorded content [11]. Although platform providers have implemented automated filtering mechanisms, gambling-related promotional content continues to appear in comment sections, suggesting limitations in detecting covert and context-dependent promotional strategies. [11]. Consequently, online comment sections have become a new avenue for covert gambling promotions, necessitating more advanced detection approaches.

The task of automatically identifying and classifying online content has been widely explored through Natural Language Processing (NLP) techniques. One of its most established applications of NLP, sentiment analysis, involves determining the emotional tone or polarity expressed in textual data. This analytical capability has evolved into broader text classification applications such as fake news detection [12], phishing email identification [13], and hate speech recognition [14]. Recent developments in transformer-based architectures, particularly Large Language Models (LLMs), have dramatically improved the performance of NLP systems across multiple domains. These models leverage contextual understanding and bidirectional attention mechanisms to capture semantic nuances in text, making them highly suitable for tasks involving complex linguistic patterns such as covert advertisements or disguised promotional content.

Several prior studies have explored the application of NLP for gambling-related analysis. Wang and Li [15] proposed a hybrid multimodal data-fusion model to identify gambling websites, achieving exceptional precision, recall, and F1-scores above 99%. In parallel, Knaebe et al. [16] utilized a LLM to detect indicators of problematic gambling behavior from online textual data. In the Indonesian research landscape, prior work has applied traditional machine learning and deep learning methods, such as Random Forest, Logistic Regression, and Convolutional Neural Networks (CNN), to detect online gambling advertisements on X (formerly Twitter) [17].

Despite these advancements, most prior works [17][18][19][20], primarily focus on detecting explicit gambling advertisements or identifying gambling-related websites using conventional machine learning or surface-level text-mining techniques. Such approaches are generally effective for clearly structured promotional content but remain limited in capturing implicit promotional patterns, slang variations, and context-dependent expressions commonly found in social media comment sections. Covert gambling promotions often employ indirect phrasing, obfuscated brand references, or informal linguistic styles that require deeper contextual modeling beyond keyword matching or shallow neural architectures. Although transformer-based models have demonstrated strong contextual representation capabilities in harmful content moderation tasks [21], their effectiveness in detecting implicit gambling discourse within Indonesian-language comments has not been systematically evaluated. This limitation highlights the need for a comparative investigation of transformer architectures in modeling nuanced and informal gambling-related expressions in Bahasa Indonesia.

Therefore, this study conducts a comparative evaluation of transformer-based language models, specifically IndoBERT and mBERT, to detect both explicit and implicit gambling-related expressions in Indonesian social media comments. To achieve this, an annotated dataset of Indonesian-language comments was constructed to support supervised model training. The proposed model was evaluated using standard text classification metrics, including precision, recall, F1-score, and accuracy, to ensure its reliability and robustness in real-world detection scenarios. By extending the application of

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

transformer-based models to comment-level detection in a low-resource linguistic context, this study advances beyond prior post-level or website-level detection approaches.

This research contributes to the expanding body of NLP literature on social media content moderation in several key dimensions. Contextually, this study extends prior gambling detection research to the Indonesian-language comment domain, addressing linguistic characteristics and informal discourse patterns specific to social media environments. Methodologically, it underscores the capability of transformer architectures to capture nuanced and informal language patterns characteristic of social media discourse. From a practical standpoint, the findings provide a foundation for developing automated moderation systems that can support social media platforms and regulatory authorities in curbing the proliferation of illegal gambling promotions within Indonesia's digital ecosystem.

2. METHOD

The overall methodological framework of this study is illustrated in Figure 1. The process begins with annotated data collected from Indonesian social media comment sections, followed by a series of preprocessing steps including data cleaning, normalization, stopword removal, and stemming to ensure text consistency and noise reduction. The cleaned dataset is then partitioned into training, validation, and test sets to support supervised learning and performance evaluation. Two transformer-based models namely IndoBERT and multilingual BERT (mBERT) that are subsequently fine-tuned for the binary classification task of detecting gambling-related comments. Each model undergoes tokenization, definition of evaluation functions, and iterative training using the preprocessed data. After training, both models are evaluated using standard classification metrics, and their performances are compared to determine the most effective architecture for identifying gambling-promotional content in Indonesian social media comments.

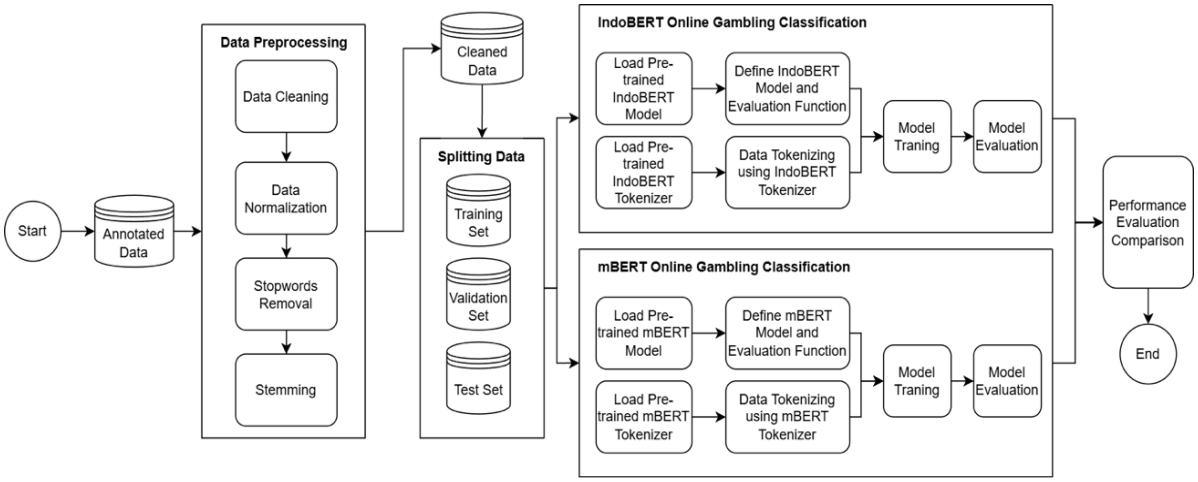

Figure 1. Architecture of Proposed Method

2.1. Data Collection

The dataset employed in this study comprises a total of 31,230 annotated comments obtained from two social media platforms: YouTube and Twitter (X). As shown in Figure 2, the dataset consists of 29,928 comments from YouTube and 1,302 comments from Twitter, encompassing both gambling-related and non-gambling content. Each record includes four attributes: user, original_message, source, and label that summarized in Table 1. The user column contains anonymized identifiers, while original_message stores the original text of the comment. The source column denotes the originating platform, and label represents the binary annotation, where 1 indicates comments related to online gambling and 0 represents unrelated content.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

The dataset was obtained from a third-party data provider and included pre-existing binary labels indicating gambling-related (1) and non-gambling (0) comments. To ensure ground-truth reliability, 10% of the dataset was randomly sampled and independently re-evaluated by two reviewers using predefined criteria for explicit and implicit gambling-promotional content. Inter-rater agreement on the sampled subset was measured using Cohen's Kappa, yielding $\kappa = 0.85$, indicating strong agreement. Minor inconsistencies identified during verification were corrected prior to model training to enhance label consistency.

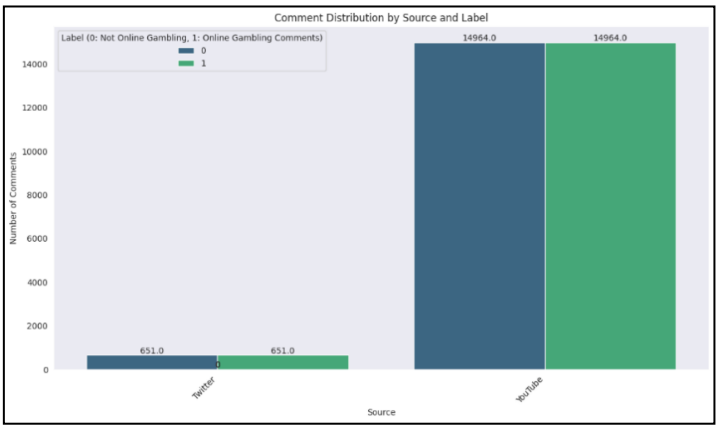

Figure 2. Comment Distribution by Source and Label

Table 1. Data Descriptions

Attribute Name	Descriptions
user	An anonymized user identifier from the social media platform
original_message	The raw text of the user comment
source	The platform where the comment originated (YouTube or Twitter)
label	Binary label indicating gambling-related content (1) or non-gambling (0)

2.2. Data Preprocessing

Data preprocessing is a crucial step in NLP that ensures textual data is clean, consistent, and ready for model training. Raw text data often contains noise such as special characters, inconsistent capitalization, and irrelevant words, which can degrade model performance. Therefore, appropriate preprocessing techniques are applied to improve both efficiency and accuracy of downstream tasks such as sentiment analysis and text classification [22][23]. The main preprocessing stages include data cleaning and normalization. Unlike traditional feature-based approaches, stemming and stopword removal were not applied, as transformer-based models rely on subword tokenization mechanisms that preserve contextual and morphological information. Table 2 shows the comment data before and after going through the preprocessing process.

Table 2. Comments Data

Original Comment	Final Comment
Bermain di *DO**RA**77* membuat saya mampu mengatasi beban hutang dengan hasil nyata.	bermain di o membuat saya mampu mengatasi beban hutang dengan hasil nyata.
Depo 100, langsung mekswin di *AERO* 8 **8** s AXL**77** gua cba bentar, eh lngsg dapet jackpot	depo 100, langsung mekswin di l gua cba bentar, eh lngsg dapet jackpot
Saya gemar dengan Matematika.... Insya Allah, tahun depan aku juga mau ambil fakultas pendidikan matematika mudah mudahan aku bisa menyusul ke Jepang Aminnn 🙏	saya gemar dengan matematika.... insya allah, tahun depan aku juga mau ambil fakultas pendidikan matematika mudah mudahan aku bisa menyusul ke jepang aminnn

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

Data cleaning involves removing unwanted elements such as URLs, punctuation, emojis, numbers, and non-standard characters. This process helps eliminate noise that could mislead the model's learning process. In sentiment and text classification studies, cleaning has been shown to significantly enhance data quality and improve classification accuracy when combined with case folding and tokenization [24]. Furthermore, recent study emphasized that the combination of cleaning and case folding yielded the highest ROUGE score in text summarization tasks, demonstrating the importance of cleaning in preserving textual integrity [25].

Data normalization standardizes textual input to reduce lexical variation and noise, improving model interpretability and performance. This process typically includes lowercasing, punctuation and symbol removal, and normalization of non-standard terms. Effective normalization ensures consistent tokenization and prevents data sparsity in embeddings. Recent work demonstrated that transformer-based normalization models significantly improve contextual representation and reduce perplexity in noisy datasets [26]. In this study, normalization was applied by converting text to lowercase, removing extraneous symbols, and standardizing special characters to produce consistent and semantically stable input for the subsequent NLP stages.

Stopwords removal eliminates high-frequency, semantically weak words (e.g., "*dan*," "*atau*," "*wkwk*") that add computational overhead and obscure informative terms. This process reduces dimensionality and enhances classification efficiency. Empirical findings confirm that tailoring stopword lists to domain-specific corpora improves text mining accuracy and interpretability [27]. Accordingly, this research applies a hybrid approach: combining standard Indonesian stopword lists with corpus-based frequency filtering to remove contextually irrelevant words while retaining domain-specific lexical items essential for accurate feature representation.

Stemming consolidates inflected or derived words into a common base form to reduce redundancy and enhance generalization in textual features. This step is particularly crucial for morphologically rich languages like Indonesian. Studies have shown that effective stemming improves classification accuracy and reduces feature sparsity [28]. Consistent with these findings, this study employs the Sastrawi stemmer, an Indonesian morphological stemmer designed to handle affixation efficiently. Its integration ensures uniform root forms across tokens, enhancing representation quality for IndoBERT and mBERT tokenization in subsequent stages.

2.3. Tokenization and Classification Using Pre-trained Models

The classification process in this study leverages pre-trained transformer-based models, specifically IndoBERT and Multilingual BERT (mBERT), to perform text understanding and sentiment classification tasks. Both models utilize the BERT architecture, which applies bidirectional self-attention to capture contextual dependencies in text [29]. IndoBERT is tailored for the Indonesian language, while mBERT provides multilingual capabilities for cross-lingual representation learning. These models have demonstrated strong generalization in low-resource languages and are effective in both monolingual and cross-lingual tasks.

Prior to model training, textual input was tokenized using the WordPiece algorithm, which segments words into subword units to handle rare and unseen tokens effectively. IndoBERT and mBERT employ different vocabularies and tokenization schemes optimized for Indonesian and multilingual corpora, respectively. In this study, tokenization was implemented through a custom function that applied truncation and padding with a maximum sequence length of 128 like shown in Table 3 about the model parameters and training configuration, ensuring consistent input dimensions across the dataset. The use of the Hugging Face tokenizer ensured compatibility with both models and reduced potential tokenization bias, particularly for Indonesian morphological variations.

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

Table 3. Model Parameters and Training Configuration

Parameter	Value
Pre-trained Model	IndoBERT, mBERT
Tokenizer	WordPiece (IndoBERTTokenizer, mBERTTokenizer)
Sequence Length	128 tokens
Optimizer	AdamW
Learning Rate	2e-5
Weight Decay	0.01
Loss Function	CrossEntropyLoss
Batch Size	16
Training/Test/Validation Split	80% / 10% / 10%

The dataset was divided into training, validation, and testing subsets using an 80-10-10 split to maintain balanced class representation. Encoded sequences were transformed into tensors and fed into PyTorch DataLoader objects for efficient batching and shuffling. This partitioning strategy ensures robust generalization and prevents overfitting by allowing model evaluation across unseen data subsets. Each encoded text sample included token IDs, attention masks, and label tensors aligned with model input specifications [30].

Model training employed AdamW optimization with a learning rate of 2e-5 and a weight decay of 0.01, consistent with best practices for fine-tuning transformer architectures. The CrossEntropyLoss function was used for multi-class classification, computed between the model's predicted logits and target labels. IndoBERT was instantiated from "indolem/indobert-base-uncased" and mBERT from "bert-base-multilingual-cased". Both models were fine-tuned using GPU acceleration to expedite computation and ensure stable convergence. Similar hyperparameter configurations have been validated in recent NLP benchmarks for Indonesian text processing [31].

In the context of the present study, with tokenization, encoding, and fine-tuning executed under matched conditions for both models, we can attribute performance differences to the influence of pre-training and vocabulary/task alignment. Results will include accuracy, precision, recall and F1-score on the held-out test set, enabling a rigorous comparison between IndoBERT and mBERT.

3. RESULT

The implementation phase involved fine-tuning two Transformer-based models: IndoBERT and multilingual BERT (mBERT). Both models were initialized with pre-trained weights—"indolem/indobert-base-uncased" and "bert-base-multilingual-cased" and extended with a simple binary classification head to distinguish between gambling-related and non-gambling comments. The models were trained using AdamW optimizer with a learning rate of 2e-5 and Cross-Entropy Loss, following standard fine-tuning procedures in low-resource text classification tasks [32]. The training utilized a batch size of 16 for 10 epochs, with early stopping behavior monitored via validation loss to mitigate overfitting.

The input texts were tokenized with a maximum sequence length of 128 to balance computational efficiency and contextual representation. Encoded inputs were batched and fed into the models using the PyTorch DataLoader. As shown in Figure 3, IndoBERT's training loss consistently decreased across epochs, while the validation loss stabilized after the 4th epoch, indicating that the model converged effectively without severe overfitting. This suggests IndoBERT's pretraining on Indonesian corpora offered a linguistic advantage for capturing contextual nuances within social media comments.

For mBERT, the Figure 4 trend shows a similar reduction in training loss but a less stable validation loss pattern, suggesting mild overfitting and lower generalization capability compared to IndoBERT. This observation aligns with prior research indicating that mBERT may underperform on monolingual Indonesian data due to its multilingual tokenization and limited exposure to informal

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

Indonesian slang [33]. Nonetheless, the model still achieved consistent convergence, demonstrating its ability to learn discriminative features for online gambling detection. Overall, the training curves indicate that IndoBERT achieved faster and more stable convergence, implying that domain-specific and language-tailored pretraining leads to superior performance in this context.

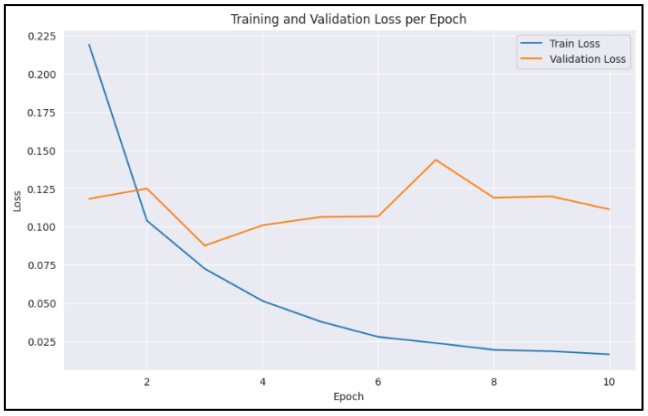

Figure 3. IndoBERT Training and Validation Loss per Epoch

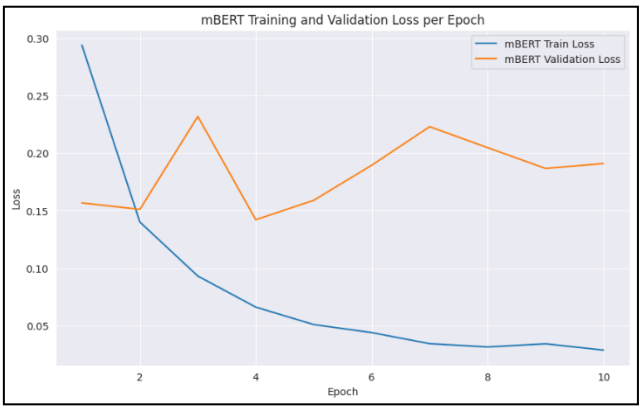

Figure 4. mBERT Training and Validation Loss per Epoch

Table 4. Quantitative Model Comparison between IndoBERT and mBERT

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT	0.98	0.98	0.97	0.98
mBERT	0.96	0.96	0.96	0.96

The performance of the IndoBERT and mBERT models was evaluated using standard classification metrics to measure their effectiveness in detecting online-gambling comments on Indonesian social-media data. The dataset used for testing consisted of 3,123 comments that were evenly distributed between the gambling (label 1) and non-gambling (label 0) classes. These evaluation metrics were selected due to their relevance in binary classification tasks, where both false positives and false negatives carry significant interpretive consequences.

Table 4 summarizes the quantitative comparison between IndoBERT and mBERT. IndoBERT achieved an overall accuracy of 0.98, surpassing mBERT's 0.96. Across all individual metrics, IndoBERT consistently demonstrated higher precision (0.97–0.98) and recall (0.97–0.98), indicating stronger sensitivity to both positive and negative classes. This improvement can be attributed to IndoBERT's language-specific pre-training on large-scale Indonesian corpora, which allows the model to capture lexical nuances, slang, and syntactic patterns typical in informal online communication. To determine whether the performance difference between IndoBERT and mBERT was statistically

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

significant, McNemar's test was applied to their paired predictions on the test set. The analysis yielded $p = 0.023$ ($p < 0.05$), indicating that the superior performance of IndoBERT is statistically significant and unlikely to be attributed to random variation.

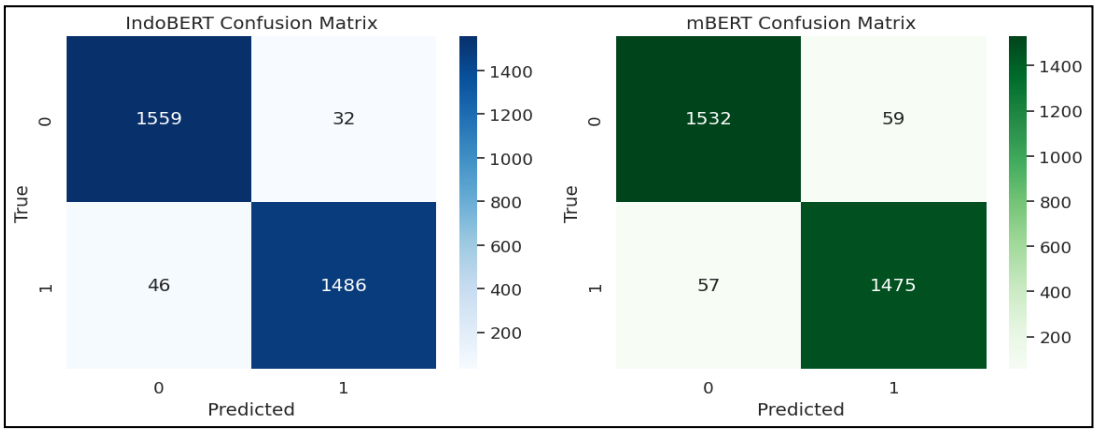

Figure 5. IndoBERT and mBERT Confusion Matrix

The confusion matrices in Figure 5 visualize the classification outcomes for each model. IndoBERT produced 1,559 true negatives and 1,486 true positives, misclassifying only 78 samples in total. In contrast, mBERT yielded 1,532 true negatives and 1,475 true positives, with 116 misclassifications. The relatively lower number of false positives and false negatives in IndoBERT demonstrates its superior balance between precision and recall.

Table 5. Quantitative Model Comparison between IndoBERT and mBERT

Original Comment	True Label	Predicted (IndoBERT)	Error Type
"main bareng biar hoki terus"	1	0	Implicit promotional phrase
"taruhan kecil aja seru kok"	0	1	Metaphorical betting expression
"gas depo tipis dulu"	1	0	Slang/abbreviation ambiguity

As shown in Table 5, misclassifications primarily occurred in comments containing implicit promotional language, metaphorical betting expressions, or domain-specific slang. For example, phrases such as "gas depo tipis dulu" rely on abbreviated gambling terminology that may not be explicitly represented in pretraining corpora. Similarly, metaphorical expressions such as "taruhan kecil" can be contextually non-gambling, leading to false positives. These findings indicate that contextual ambiguity and informal slang remain key challenges even for transformer-based architectures.The evaluation also demonstrates that both models achieved relatively high performance despite differences in linguistic specialization. While mBERT's multilingual nature enables broad generalization across languages, its tokenization granularity limits deep understanding of language-specific morphology, which is crucial for short informal comments. IndoBERT's ability to maintain higher recall while preserving precision highlights its suitability for real-world monitoring systems where missing actual gambling-related content could pose regulatory risks.

Although both IndoBERT and mBERT achieved high accuracy scores, 0.98 and 0.96 respectively, the confusion matrices in Figure 5 reveal several misclassifications that merit closer examination. The majority of the errors for both models occurred in differentiating borderline comments containing ambiguous or euphemistic language. For instance, comments using indirect promotional phrases such as "*main bareng biar hoki*" ("play together for luck") or coded symbols (e.g., numbers or emojis representing betting platforms) were occasionally misclassified as non-gambling content. This indicates

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

that even advanced Transformer models can struggle to interpret implicit semantic cues in informal Indonesian social media discourse.

IndoBERT's language-specific pretraining contributed to fewer false positives (32) compared to mBERT (59), suggesting stronger contextual grounding in Indonesian syntax and idiomatic expressions. Conversely, mBERT produced slightly more balanced misclassification patterns across both classes, reflecting its multilingual exposure but also its lesser sensitivity to localized slang. Several false negatives observed in both models were linked to short or fragmented comments lacking clear linguistic indicators of gambling activity, emphasizing the challenge of low-context data. Overall, the error distribution highlights the importance of cultural and linguistic adaptation in fine-tuning Transformer models for Indonesian data. Future work could explore integrating domain-specific lexicons or leveraging multimodal features, such as emojis and hashtags to reduce ambiguity. Additionally, incorporating adversarial training or active learning strategies may further enhance robustness against evolving linguistic patterns used by online gambling promoters.

To further assess the effectiveness of Transformer-based architectures, an additional experiment was conducted using a Long Short-Term Memory (LSTM) network as a baseline model. The LSTM was trained on the same dataset with identical preprocessing, using word embeddings to capture sequential linguistic features. As shown in the evaluation, the LSTM model achieved an accuracy of 91%, with an F1-score of 0.91 for both classes. This indicates that while the LSTM was able to capture temporal dependencies in the data, its performance remained lower than that of the Transformer-based models.

The superior results achieved by IndoBERT (F1-score 0.98) and mBERT (F1-score 0.96) highlight the advantage of bidirectional attention mechanisms and contextualized word representations over sequential encoding. Unlike LSTM, which processes text in a fixed order, Transformer models leverage self-attention to model long-range dependencies more effectively, particularly in informal and code-mixed Indonesian comments. Therefore, these findings confirm that the adoption of Transformer-based models, especially language-specific ones such as IndoBERT, provides a more accurate and context-aware approach for detecting online gambling content on social media platforms.

In addition to the LSTM baseline, a traditional machine learning classifier using a Linear Support Vector Machine (SVM) with TF-IDF features was implemented to provide a lightweight benchmark. The SVM model achieved an accuracy of 0.88 and an F1-score of 0.86, which is substantially lower than the Transformer-based models. While SVM offers lower computational complexity and faster training time, its reliance on surface-level lexical features limits its ability to capture contextual and implicit promotional patterns in informal Indonesian comments. This comparison highlights the trade-off between computational efficiency and contextual modeling capability.

4. DISCUSSIONS

The experimental results demonstrate a clear performance hierarchy across modeling paradigms. IndoBERT achieved the highest F1-score of 0.98, followed by mBERT at 0.96, LSTM at 0.91, and Linear SVM at 0.88. This progression reflects the increasing representational depth of the respective architectures. Traditional machine learning models such as SVM rely on sparse lexical representations, which restrict their ability to capture contextual nuance. Sequential neural networks such as LSTM improve contextual modeling but remain constrained in handling long-range semantic dependencies. Transformer-based architectures, grounded in self-attention mechanisms as introduced by Devlin et al. [29], provide richer contextual encoding that enables the detection of implicit promotional cues and subtle linguistic patterns frequently observed in gambling-related discourse.

When compared to prior Indonesian studies on gambling promotion detection, the performance of the proposed approach demonstrates measurable advancement. Perdana et al. [17] applied text mining

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

algorithms for detecting online gambling promotions on Indonesian Twitter, reporting competitive performance using classical and hybrid approaches. However, their study primarily relied on conventional feature engineering strategies, which are inherently limited in capturing covert and context-dependent expressions. Similarly, Kamdan et al. [18] evaluated IndoBERT for detecting gambling promotion in YouTube comments, yet their reported results were slightly lower than the F1-score achieved in the present study. The improvement observed here may be attributed to dataset scale, refined fine-tuning procedures, and stricter annotation validation. This indicates that performance optimization is not solely dependent on architecture choice but also on data quality and annotation reliability.

In addition, Maldini et al. [19] proposed a multimodal framework integrating Faster R-CNN and Tr-OCR for detecting covert gambling advertisements in visual formats. While their approach extends beyond textual classification and addresses image-based promotions, it operates in a different modality context. The present study complements such multimodal research by demonstrating that text-only Transformer models can already achieve near-saturated performance levels in purely linguistic settings. This comparison highlights that while multimodal integration may enhance robustness, high-quality monolingual contextual modeling remains foundational for textual moderation pipelines.

The comparison between IndoBERT and mBERT further reinforces the importance of linguistic specialization. Although both models share the same Transformer backbone, the 2 percent F1 difference is statistically significant with a p-value of 0.023. This confirms that the improvement is not incidental. Dhendra and Utomo [31] similarly reported that IndoBERT outperforms general Transformer variants in Indonesian sentiment classification tasks, emphasizing the benefit of monolingual pretraining. The current findings extend this evidence to sensitive content detection, demonstrating that language-aligned pretraining enhances the model's ability to interpret informal syntax, slang, and culturally embedded promotional strategies.

The complexity of informal Indonesian discourse also aligns with findings in sarcasm detection research. Fitrianto et al. [14] showed that detecting Indonesian sarcasm requires deeper contextual modeling due to implicit meaning and pragmatic cues. Error analysis in the present study reveals that most misclassifications occur in comments containing metaphor, humor, or indirect references to gambling platforms. This similarity suggests that gambling promotion detection shares characteristics with sarcasm and malicious intent detection, where meaning is often implied rather than explicitly stated. Consequently, contextual self-attention mechanisms play a critical role in distinguishing subtle semantic signals.

Beyond performance comparison, computational considerations remain essential for deployment feasibility. Fine-tuning IndoBERT requires substantially higher GPU memory allocation and longer training time compared to LSTM and SVM models. Although the 2 percent improvement over mBERT is statistically significant, stakeholders must evaluate whether this marginal gain justifies additional computational cost in large-scale moderation systems. In high-volume social media environments, latency and scalability become operational priorities. Therefore, mBERT may serve as a computationally efficient alternative in infrastructure-constrained settings, while IndoBERT remains preferable when maximizing detection precision is critical.

Another important dimension concerns the ethical implications of False Positive predictions. In automated moderation systems, incorrectly flagging legitimate user comments as gambling-related may lead to unjustified content removal and reduced platform trust. Even with high precision, small false positive rates can scale into substantial moderation errors when applied to millions of comments. Therefore, deployment should incorporate confidence-based thresholds, appeal mechanisms, or human verification layers to ensure balanced enforcement. This consideration aligns with broader discussions on responsible AI governance in sensitive content detection systems.

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

Overall, the findings position IndoBERT-based modeling as a robust and empirically validated solution for detecting gambling-related promotional content in Indonesian social media. Compared with prior text-based and multimodal studies, the present work contributes by providing statistically validated improvements, computational trade-off analysis, and ethical deployment considerations. The results reinforce the strategic importance of language-specific Transformer pretraining in addressing localized digital moderation challenges.

5. CONCLUSION

This study confirms that monolingual Transformer models, particularly IndoBERT, provide superior performance for detecting gambling-related promotional comments in Indonesian social media. The statistically significant improvement over multilingual and sequential baselines indicates that language-specific pretraining enhances contextual sensitivity to informal expressions, implicit persuasion, and localized sociolinguistic patterns. These findings suggest that effective moderation of sensitive digital content requires linguistic alignment between pretrained models and target discourse environments.

Beyond performance metrics, the study demonstrates that contextual self-attention architectures are structurally better suited for identifying covert promotional language compared to surface-based or sequential approaches. However, model selection should also consider computational trade-offs, as higher accuracy is accompanied by increased training cost and infrastructure requirements.

This research is limited by its focus on two platforms and text-only data, which may not fully represent the multimodal nature of contemporary gambling promotion strategies. Future work should explore multimodal detection frameworks integrating textual and visual signals, as well as lightweight Transformer optimization techniques to improve scalability for real-time moderation systems. These developments are essential for building accurate, efficient, and responsible automated moderation infrastructures in localized digital ecosystems.

REFERENCES

[1] A. S. Wardani, "Makin Gawat, 960 Ribu Pelajar dan Mahasiswa Terjerat Judi Online," Liputan6.com. Accessed: Oct. 06, 2025. [Online]. Available: https://www.liputan6.com/tekno/read/5799064/makin-gawat-960-ribu-pelajar-dan-mahasiswa-terjerat-judi-online

[2] N. Azka Syakira, N. Fathma Ramadhahana, N. Devi Anggita, T. Tsaqifa, and R. Naila Husna, "Dampak Konsumerisme Berupa Judi Online di Indonesia: Perspektif Ekonomi, Sosial, dan Mental," *Jurnal Interaktif*, vol. 16, no. 2, pp. 73–79, Dec. 2024, doi: 10.21776/ub.interaktif.2024.016.02.3.

[3] S. Sriyana, "JUDI ONLINE: DAMPAK SOSIAL, EKONOMI, DAN PSIKOLOGIS DI ERA DIGITAL," *JURNAL SOCIOPOLITICO*, vol. 7, no. 1, pp. 27–34, Feb. 2025, doi: 10.54683/sociopolitico.v7i1.169.

[4] A. Price, "Online Gambling in the Midst of COVID-19: A Nexus of Mental Health Concerns, Substance Use and Financial Stress," *International Journal of Mental Health and Addiction*, vol. 20, no. 1, pp. 362–379, Feb. 2022, doi: 10.1007/s11469-020-00366-1.

[5] A. Sirola, N. Savela, I. Savolainen, M. Kaakinen, and A. Oksanen, "The Role of Virtual Communities in Gambling and Gaming Behaviors: A Systematic Review," *J Gambl Stud*, vol. 37, no. 1, pp. 165–187, Mar. 2021, doi: 10.1007/s10899-020-09946-1.

[6] R. Suriá-Martínez, F. García-Castillo, E. Villegas-Castrillo, C. López-Sánchez, and C. Carretón-Ballester, "Negative impact of online gambling problematic in disabled and non-disabled university students: exploring the risk profile," *Front Psychol*, vol. 15, Sep. 2024, doi: 10.3389/fpsyg.2024.1429122.

[7] Populix, "Understanding the impact of online gambling ads exposure," Populix. [Online]. Available: https://info.populix.co/product/consumer-trend-report

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

[8] E. Bolat, C. Panourgia, A. Yankouskaya, and M. Kelly, "Influencer-Driven Gambling Content and Its Impact on Children and Young People: A Scoping Study," *Curr Addict Rep*, vol. 12, no. 1, p. 3, Jan. 2025, doi: 10.1007/s40429-025-00616-z.

[9] Sudirham and T. B. Sari, "Adapting counter-gambling advertising to the Indonesian context: a call to action," *J Public Health (Bangkok)*, vol. 47, no. 2, pp. e246–e247, May 2025, doi: 10.1093/pubmed/fdae221.

[10] Z. K. Muharam, W. Astuti, R. Prasida, and D. Syahputra, "Indonesian Journal of Digital Public Relations (IJDPR) PENGGUNAAN INFLUENCER DALAM PROMOSI JUDI ONLINE DAN SENTIMEN PUBLIK THE USE OF INFLUENCERS IN PROMOTION ONLINE GAMBLING AND PUBLIC SENTIMENT," 2024. [Online]. Available: https://journals.telkomuniversity.ac.id/IJDPR

[11] Kumparan News, "Marak iklan judol di YouTube, apa kata Google Indonesia?," Kumparan. [Online]. Available: https://kumparan.com/kumparannews/marak-iklan-judol-di-youtube-apa-kata-google-indonesia-24WaAnoJ08G/

[12] N. C. Harriott and A. L. Ryan, "Proteomic profiling identifies biomarkers of COVID-19 severity," *Heliyon*, vol. 10, no. 1, p. e23320, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23320.

[13] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nuñez-Agurto, and G. Rodríguez-Galán, "A Phishing-Attack-Detection Model Using Natural Language Processing and Deep Learning," *Applied Sciences*, vol. 13, no. 9, p. 5275, Apr. 2023, doi: 10.3390/app13095275.

[14] R. A. Fitrianto, A. S. Editya, M. M. Alamin, A. L. Pramana, and A. K. Alhaq, "Classification of Indonesian Sarcasm Tweets on X Platform Using Deep Learning," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, Jul. 2024, pp. 388–393. doi: 10.1109/ICICoS62600.2024.10636904.

[15] S. Oh, K. Chung, and J. Choi, "Resource-Oriented Augmentation of a Train Timetable," *IEEE Access*, vol. 11, pp. 114283–114290, 2023, doi: 10.1109/ACCESS.2023.3323590.

[16] M. Y. Ali, A. M. Yimer, and T. S. Dessie, "An empirical estimation of aggregate import demand under foreign exchange constraints: Evidence from Ethiopia," *PLoS One*, vol. 19, no. 6, p. e0303587, Jun. 2024, doi: 10.1371/journal.pone.0303587.

[17] R. B. Perdana, A. -, I. Budi, A. B. Santoso, A. Ramadiah, and P. K. Putra, "Detecting Online Gambling Promotions on Indonesian Twitter Using Text Mining Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 8, 2024, doi: 10.14569/IJACSA.2024.0150893.

[18] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments," in *The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society*, Basel Switzerland: MDPI, Sep. 2025, p. 66. doi: 10.3390/engproc2025107066.

[19] A. S. Maldini, W. S. J. Saputra, and D. A. Prasetya, "Multimodal Detection of Covert Online Gambling Advertisements Using Faster R-CNN and Tr-OCR," *bit-Tech*, vol. 8, no. 1, pp. 953–963, Aug. 2025, doi: 10.32877/bt.v8i1.2769.

[20] K. A. Adriana and E. B. Setiawan, "Enhancing Cyberbullying Detection with a CNN-GRU Hybrid Model, Word2Vec, and Attention Mechanism," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 3, pp. 1113–1130, Jun. 2025, doi: 10.52436/1.jutif.2025.6.3.4176.

[21] Z. Fang, H. Zhang, J. He, Z. Qi, and H. Zheng, "Semantic and Contextual Modeling for Malicious Comment Detection with BERT-BiLSTM," arXiv preprint arXiv:2503.11084, 2025. doi: 10.48550/arXiv.2503.11084.

[22] M. F. Cahyadi and T. H. Rochadiani, "Implementasi Ensemble Deep Learning Untuk Analisis Sentimen Terhadap Genre Game Mobile," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 3, p. 1512, Jul. 2024, doi: 10.30865/mib.v8i3.7832.

[23] Teddy Oswari, Murniyati, Trityanti Yusnitasari, Nurasiah, and Seviyanti Wijay, "Sentiment Analysis of Indonesian YouTube Reviews About Lesbian, Gay, Bisexual, and Transgender (LGBT) using IndoBERT Fine Tuning," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 15, no. 01, pp. 26–37, Oct. 2025, doi: 10.24843/LKJITI.2024.v15.i01.p03.

Vol. 7, No. 2, April 2026, Page. 1931-1943
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5677

[24] Sofyan Hidayat, Nining Rahaningsih, Raditya Danar Dana, and Mulyawan, "Improvement of User Sentiment Classification Model for the Indomaret Poinku Application Using the Naïve Bayes Method," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 2, pp. 1497–1500, Feb. 2025, doi: 10.59934/jaiea.v4i2.937.

[25] M. F. Juna and M. Hayaty, "The observed preprocessing strategies for doing automatic text summarizing," *Computer Science and Information Technologies*, vol. 4, no. 2, pp. 119–126, Jul. 2023, doi: 10.11591/csit.v4i2.p119-126.

[26] M. Ashmawy, M. W. Fakhr, and F. A. Maghraby, "Lexical Normalization Using Generative Transformer Model (LN-GTM)," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 183, Nov. 2023, doi: 10.1007/s44196-023-00366-8.

[27] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS One*, vol. 16, no. 8, p. e0254937, Aug. 2021, doi: 10.1371/journal.pone.0254937.

[28] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," Journal of Big Data, vol. 8, no. 1, p. 26, Dec. 2021, doi: 10.1186/s40537-021-00413-1.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[30] N. Surantha, P. Atmaja, David, and M. Wicaksono, "A Review of Wearable Internet-of-Things Device for Healthcare," *Procedia Computer Science*, vol. 179, pp. 936–943, 2021, doi: 10.1016/j.procs.2021.01.083.

[31] Dhendra and V. Gayuh Utomo, "Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews," *Jurnal Transformatika*, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019.

[33] Z. Lin *et al.*, "Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue StateTracking," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 5640–5648. doi: 10.18653/v1/2021.naacl-main.448.