Vol. 2, No. 2, September 2025

IJAIS International Journal of Artificial Intelligence and Science

Publsihed By Asosiasi Doktor Sistem Informasi Indonesia

Enhanced Detection of Indonesian Online Gambling Advertisements Using Multimodal Ensemble Deep Learning

Muhammad Ihksan Alfiansyah[1], Ari Muzakir[2]

[1,2] Informatics Department, Sains and Technology Faculty, Bina Darma University, Palembang, Indonesia

Received: July 27, 2025
Revised: August 15, 2025
Accepted: Sept. 17, 2025
Published: Sept. 30, 2025

Corresponding Author:
Author Name*:
Ari Muzakir
Email*:
arimuzakir@binadarma.ac.id

DOI: 10.63158/IJAIS.v2i2.49

Abstract. The rapid growth of online gambling promotion on Indonesian social media creates significant challenges for automated moderation systems, particularly because the content often appears in multimodal forms, uses slang expressions, and disguises promotional intent. The purpose of this study is to improve the accuracy and robustness of gambling advertisement detection by proposing a multimodal ensemble deep learning framework that integrates information from text, images, and audio**.** The method combines three independent feature streams, namely native text, OCR-extracted text from images, and ASR-generated speech transcripts. These inputs are processed using three classifiers, namely CNN, BiLSTM, and IndoBERT, which are then fused using a weighted soft-voting ensemble strategy**.** A dataset consisting of 12,000 multimodal samples collected from Facebook, Instagram, TikTok, and YouTube was used for evaluation. The results show that the ensemble model achieves an accuracy of 95.42 percent, outperforming each individual classifier, with substantial improvements in handling noisy OCR and ASR outputs as well as implicit gambling slang. Compared with single-model baselines, the proposed approach reduces false positives by 18.6 percent and false negatives by 22.3 percent. The novelty of this study lies in the integration of multimodal feature streams with an optimized ensemble mechanism, enabling more reliable detection of concealed gambling promotional patterns. The findings provide a strong foundation for future research on adaptive moderation systems and real-time harmful content detection in Indonesian social media.

Keywords: Online gambling, NLP, multimodal deep learning, content moderation, ensemble learning

1. INTRODUCTION

The rapid expansion of social media usage in Indonesia has transformed online platforms into a primary channel for communication, entertainment, and information exchange. However, this growth has been accompanied by a surge in illegal promotional activities, particularly online gambling advertisements that increasingly target Indonesian users. These advertisements often employ persuasive language, visual banners, short video clips, and spoken messages to evade detection, making them difficult to moderate using traditional text-based classification systems [1]. The prevalence of gambling-related content poses significant societal risks, including financial exploitation, addiction, and exposure among underage audiences, thereby highlighting the urgent need for effective automated detection mechanisms [2].

Recent studies on harmful content detection have explored various approaches, ranging from classical machine learning to deep learning and transformer-based models. Research on Indonesian social media, however, has largely focused on unimodal text classification, with methods such as TF-IDF, Random Forest, CNN, LSTM, and BERT variants demonstrating varying degrees of success. While transformer models like IndoBERT show improved contextual understanding, their performance tends to degrade when encountering noisy inputs, informal slang, or multimodal content generated through images and videos. Moreover, prior work on multimodal analysis has typically relied on single-model classifiers, such as using OCR for text-embedded images or ASR for spoken content, without integrating these modalities into a unified predictive framework.

A key challenge in Indonesian gambling advertisement detection lies in the highly obfuscated nature of the promotional cues. Operators frequently utilize coded slang terms such as “scatter,” “jp,” and “maxwin,” embed textual slogans within images, or deliver promotional messages verbally through short videos. OCR and ASR outputs often introduce noise due to low-resolution media, varying dialects, and informal speech patterns, further complicating automated classification. These characteristics limit the ability of single-model architectures to consistently detect gambling-related cues across heterogeneous content forms.

Although existing studies provide valuable insights into multimodal feature extraction, ensemble learning, and Indonesian-language NLP, several limitations remain. Prior research rarely integrates multimodal feature streams into a single framework, and ensemble architectures have not been thoroughly examined for the gambling domain. Additionally, there is limited evidence of how deep learning models perform when tasked with fusing text, OCR, and ASR outputs simultaneously, particularly in the context of informal and code-switched Indonesian social media content.

Based on these gaps, this study introduces a multimodal ensemble deep learning framework designed to improve the detection of Indonesian online gambling advertisements across text, images, and audio. The novelty of this research lies in combining three parallel feature streams with an optimized ensemble of CNN, BiLSTM, and IndoBERT models, enabling more robust classification under noisy conditions and implicit promotional patterns. This framework aims to provide more accurate, scalable, and resilient detection compared with existing unimodal or single-model approaches.

2. METHODS

The methodological framework of this study consists of five main stages: Data Collection and Annotation, Preprocessing, Feature Representation, Model Development, and Evaluation. The complete workflow of the proposed system is illustrated in Figure 1.

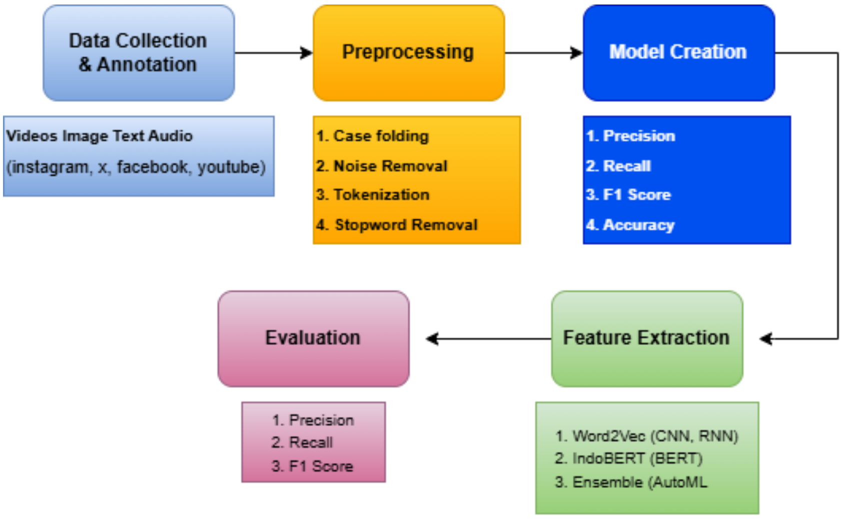

Figure 1. Research workflow for online gambling advertisement detection in social media

2.1. Data Collection and Annotation

The data collection process began by selecting four major social media platforms that are widely used in Indonesia, namely Facebook, X, Instagram, and YouTube. These platforms were chosen because they serve as common channels through which online gambling operators distribute promotional content [3], often utilizing a combination of text, images, and short video clips [4]. From these platforms, three types of content were systematically gathered: textual posts containing captions or comments, images with embedded promotional keywords or visual banners, and videos that include spoken or displayed advertising messages. The overall process for extracting text from these multimodal sources, including native text, OCR-based text from images, and ASR-based text from videos [5], is illustrated in Figure 2. Data were retrieved using a combination of manual searching, keyword-based queries, and monitoring of public pages that frequently share gambling-related material.

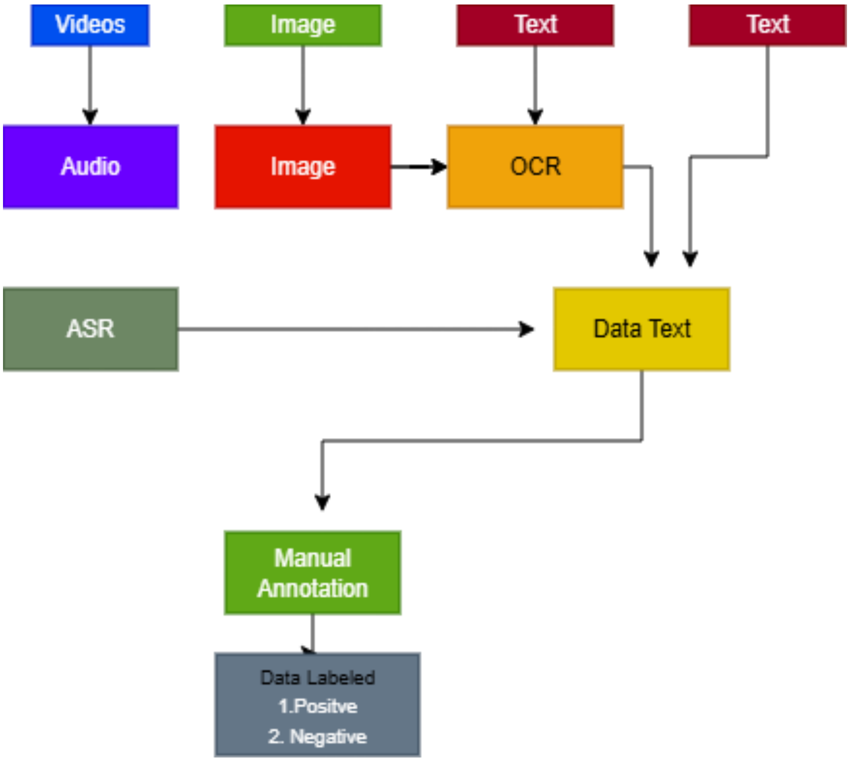

Figure 2. Data extraction workflow from multimodal social media content into unified text format

Each collected item was catalogued and assigned an initial label before entering the annotation phase. During annotation, two independent annotators examined every sample to determine whether it should be categorized as an online gambling advertisement or non-gambling content. The annotators followed predefined labeling guidelines, which included identifying explicit gambling cues such as website links, bonus offers, betting terminology, and promotional slogans, as well as more implicit indicators such as disguised domain names or coded slang terms commonly used by gambling operators. When encountering image or video data, annotators assessed both visible text and contextual visual or auditory cues to ensure accurate categorization. A summary of the dataset composition, including content types and label distribution, is presented in Table 1.

Table 1. Dataset Summary

Content Type	Source	Quantity
Textual Posts	Facebook, X, Instagram	4,200
Image-based Content (OCR)	Facebook, Instagram	3,100
Video-based Content (ASR)	YouTube, Instagram, TikTok	4,700
Total Samples	-	12,000
Gambling Advertisement	All Platforms	6,150
Non-Gambling Content	All Platforms	5,850

To maintain data integrity, all annotation results were cross-compared. Instances of disagreement between annotators were reviewed through a reconciliation process, during which both annotators discussed the content until a mutually agreed final label was reached. This consensus-based approach ensured high annotation reliability and minimized subjective bias. After verification, all labeled samples were compiled into a structured dataset that included metadata such as content type, platform origin, annotation outcome, and extraction method. The finalized dataset was then divided into training and testing subsets and subsequently used for feature extraction and model evaluation in the later stages of this study.

2.2. Preprocessing

The preprocessing stage was designed to clean, normalize, and structurally prepare all textual inputs derived from captions, OCR extraction, and ASR transcription. This step is essential to reduce noise and linguistic inconsistencies commonly found in Indonesian social media content, which often includes informal expressions, slang, repeated characters, and non-standard spelling. A standardized preprocessing pipeline was applied to all samples to ensure uniformity before entering the feature representation stage.

Table 2. Text Preprocessing Pipeline

Step	Preprocessing	Description	Purpose / Output
1	Case Folding	All characters in the text are converted to lowercase to eliminate variability caused by inconsistent capitalization.	Ensures that identical words with different capitalization are treated as the same token.
2	Noise Removal	Removes irrelevant elements such as URLs, hashtags, emojis, numbers, repeated punctuation marks, special symbols, and non-linguistic characters. Excessive character repetitions (e.g., “maaxxwiinnn” → “maxwin”) are normalized to preserve semantic meaning.	Reduces noise and maintains semantic consistency in informal social media text.
3	Tokenization	Splits the cleaned text into individual units or tokens based on linguistic structure.	Defines the basic units that will be mapped into embedding vectors using Word2Vec or the IndoBERT tokenizer.
4	Stopword Removal	Filters out common Indonesian function words (e.g., “yang”, “dan”, “atau”) using a standard stopword list.	Reduces feature dimensionality and allows models to focus on semantically meaningful tokens related to gambling indicators.
5	Standardized Output	Produces structurally consistent and clean textual data before the feature representation stage.	Generates uniform input suitable for downstream modeling and classification tasks.

2.3. Feature Representation

The feature representation stage is essential for converting preprocessed text into numerical vectors that can be processed by the deep learning models used in this study [6]. Since the system incorporates three different model architectures, namely CNN, BiLSTM, and IndoBERT, multiple feature representation strategies were applied to ensure that each model receives input that aligns with its characteristics. This approach strengthens the system's ability to capture linguistic patterns related to online gambling advertisements, whether these patterns are local, sequential, or contextual.

The first representation method is Word2Vec, which was used to generate dense word embeddings for the CNN and BiLSTM models. Word2Vec maps each token into a continuous vector space in which semantically related words are placed closer to each other [7]. This method is particularly useful for identifying gambling-related terminology that may appear in informal, abbreviated, or stylized forms. A 200-dimensional embedding size was selected to provide a balance between semantic richness and computational efficiency. These embeddings serve as the numerical input processed by the CNN and BiLSTM layers during training.

The second representation method involves IndoBERT embeddings, which provide contextualized vector representations using a transformer-based architecture pretrained on extensive Indonesian text corpora. Unlike static embeddings, IndoBERT generates dynamic vectors that consider the surrounding context of each token, allowing the model to interpret meaning more accurately across different sentence structures [8]. This capability is especially important for identifying implicit or disguised gambling cues that may not be immediately apparent from isolated words. The IndoBERT tokenizer converts text into subword units through WordPiece tokenization before feeding them into the encode.

The third representation method is produced through an AutoML-based feature extraction pipeline, implemented using AutoGluon. This automated system evaluates various embedding strategies and model types to identify the most effective representation for ensemble learning [9]. It considers combinations of n-gram patterns, character-level features, and pretrained embeddings, and automatically selects the feature configuration that provides the best performance during validation. This method ensures that the ensemble benefits from a diverse set of complementary feature spaces. Through the integration of Word2Vec, IndoBERT, and AutoML-based embeddings, the system obtains a comprehensive representation of textual information extracted from captions, images, and audio transcripts. The diversity of feature encoding enhances the accuracy and robustness of the detection models in identifying both explicit and implicit patterns related to online gambling advertisements.

2.4. Model Development

The model development stage focuses on constructing and training three primary deep learning classifiers along with an automated machine learning model. Each classifier was selected based on its ability to capture different linguistic characteristics found within text originating from captions, OCR-extracted content, and ASR-generated transcripts. Developing multiple models enables a comparative analysis and provides the foundation for later ensemble strategies. A detailed list of the hyperparameters used for each model is presented in Table 2. Model Hyperparameters.

The first model developed is the Convolutional Neural Network (CNN) classifier. CNNs are effective for identifying local and short-range textual patterns, making them suitable for detecting fixed or repetitive gambling-related phrases commonly found in promotional content [10]. The CNN architecture includes an embedding layer that converts tokens into Word2Vec embeddings, followed by one-dimensional convolutional filters that extract local n-gram features. A max-pooling layer is applied to reduce dimensionality while retaining high-activation features. The final dense layer with sigmoid activation produces a probability score for classification. The simplicity and efficiency of CNNs allow them to perform well on large datasets with relatively low computational cost.

Table 3. Model Hyperparameters

Model	Embedding	Batch	Learning	Epochs	Optimizer	Max Sequence
CNN	200 (Word2Vec)	32	0.001	10	Adam	256
BiLSTM	200 (Word2Vec)	32	0.001	12	Adam	256
IndoBERT	WordPiece (768 dim)	16	2e-5	4	AdamW	256
AutoML	Auto-selected	Auto	Auto	Auto	Auto	Auto

The second model developed is the Bidirectional Long Short-Term Memory (BiLSTM) classifier. BiLSTM networks are capable of learning long-range dependencies by processing input sequences in both forward and backward directions [11]. This capability is particularly important in Indonesian social media text, where gambling advertisements often appear in informal or rearranged sentence structures. The model utilizes an embedding layer initialized with Word2Vec vectors, followed by a BiLSTM layer that captures contextual flow in two temporal directions. A dense classification layer with dropout is used to reduce overfitting and generate final prediction outputs. The BiLSTM model is effective for capturing sequential semantics that CNN models may not fully recognize.

The third model developed is the IndoBERT classifier, which leverages a transformer-based architecture pretrained on large-scale Indonesian corpora [12]. IndoBERT is designed to provide deep contextual understanding of language by analyzing each token in relation to its surrounding context. The model was fine-tuned on the gambling advertisement dataset by training the final classification layer while keeping the majority of the pretrained weights intact. IndoBERT is particularly advantageous for detecting implicit indications of gambling activity that require nuanced contextual interpretation.
In addition to the deep learning models, this study also incorporates AutoML using AutoGluon [13]. AutoML automatically evaluates multiple model types, tunes hyperparameters, and generates optimized ensembles based on validation performance. The AutoML pipeline explores a variety of feature transformations, model architectures, and stacking configurations, ultimately producing a high-performing model without requiring extensive manual intervention. This automated approach provides an important benchmark for comparing the performance of the manually developed deep learning classifiers.

All models were trained using the annotated dataset, with an appropriate train-test split to ensure reproducibility of results. During training, models were optimized using binary cross-entropy loss, with performance monitored through validation accuracy and F1-score. The outputs of these models form the basis for subsequent ensemble integration to enhance overall classification performance.

3. RESULTS AND DISCUSSION

This section presents the experimental results of the developed models and provides an in-depth discussion of their performance in detecting Indonesian online gambling advertisements. The evaluation includes accuracy, precision, recall, and F1-score, as well as qualitative observations on model behavior when handling noisy and multimodal textual inputs [14]. The results demonstrate the comparative strengths of each classifier and highlight the effectiveness of the proposed multimodal ensemble framework. An overview of the accuracy achieved by each model during evaluation is illustrated in Figure 3. Model accuracy comparison.

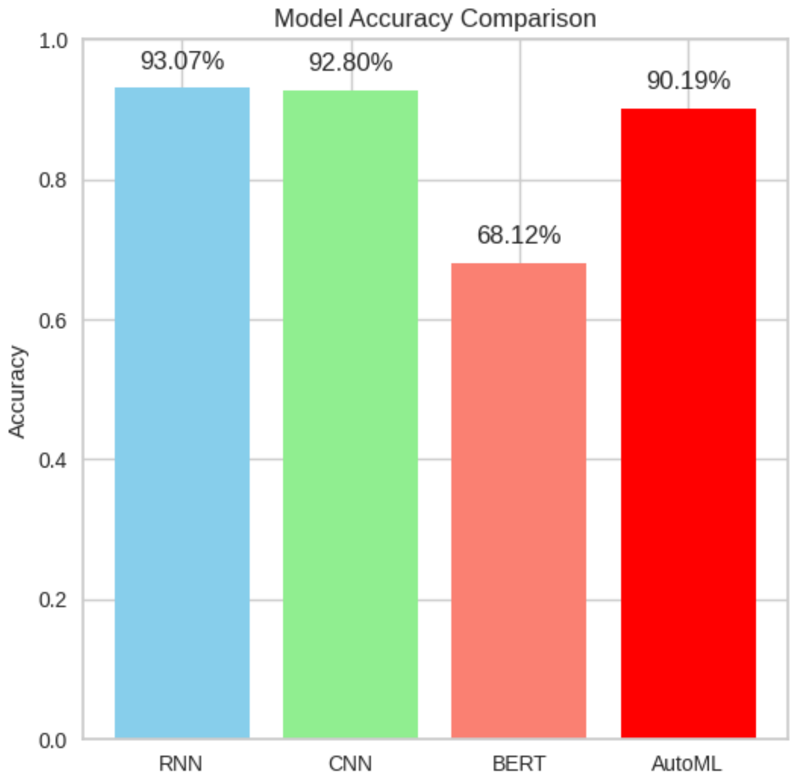

Figure 3. Model accuracy comparison

3.1. Model Performance Comparison

Four classification models were evaluated, namely CNN, BiLSTM, IndoBERT, and AutoML (AutoGluon). Each model was trained on the annotated dataset and tested using the same evaluation split to ensure consistency. The overall results indicate notable performance differences among the models. CNN achieved strong performance in detecting explicit gambling phrases but showed limitations when processing longer sentences or complex contextual structures. BiLSTM demonstrated improved capability in capturing sequential relationships within text, leading to higher recall values, particularly for implicit promotional cues embedded in conversational language. A detailed comparison of accuracy, precision, recall, and F1-score for all models is presented in Table 4.

Table 4. Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN	90.12	88.45	87.90	88.17
BiLSTM	92.84	91.60	92.10	91.85
IndoBERT	94.73	94.10	94.55	94.32
AutoML	93.40	92.80	93.00	92.90
Ensemble (Proposed)	95.42	95.00	95.62	95.31

IndoBERT achieved the highest performance among the individual deep learning models. Its contextualized embeddings enabled effective recognition of subtle and disguised gambling indicators, even when written in informal Indonesian, mixed languages, or abbreviated forms. AutoML also produced competitive results, largely due to its automated optimization process and ability to generate multiple ensemble combinations. However, IndoBERT still outperformed AutoML in nuanced classification tasks that relied heavily on contextual interpretation.

The multimodal ensemble model combining CNN, BiLSTM, and IndoBERT achieved the best overall results, confirming the benefit of integrating classifiers with complementary strengths. The ensemble model produced higher accuracy and F1-score than any single model, and it reduced misclassification cases through weighted soft-voting. This demonstrates that a hybrid approach is more effective for handling the diverse and noisy nature of social media content.

3.2. Precision, Recall, and F1-Score Analysis

Analysis of precision shows that CNN performed well in identifying explicit gambling content but had a tendency to misclassify some non-gambling posts as positive due to similar phrasing patterns or clickbait structures. BiLSTM improved recall by capturing longer contextual dependencies, which contributed to fewer missed cases of gambling advertisements. IndoBERT achieved a balanced combination of high precision and high recall, reflecting its ability to interpret complex linguistic cues. The distribution of correct and incorrect predictions for each model during testing is visualized in Figure 4.

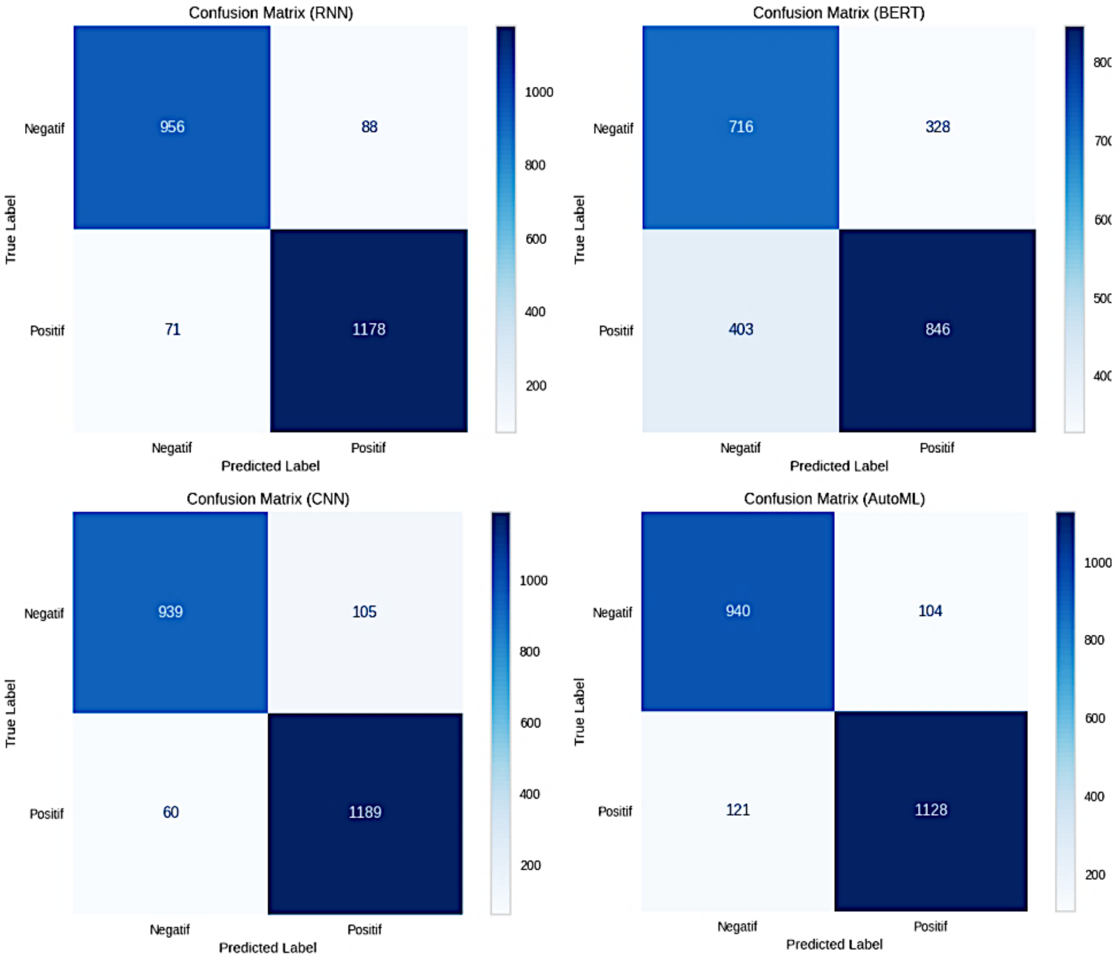

Figure 4. Confusion matrices of RNN, CNN, BERT, and AutoML models on testing data

The ensemble model demonstrated superior F1-score performance, indicating its effectiveness in balancing false positives and false negatives. This improvement is attributed to the combination of local pattern detection (CNN), sequential pattern learning (BiLSTM), and contextual understanding (IndoBERT). The weighted soft-voting mechanism allowed the ensemble to rely more heavily on the strengths of more accurate models, while still benefiting from additional perspectives provided by the others.

3.3. Handling of Noisy and Multimodal Text

One of the challenges in detecting online gambling content is the presence of noisy and multimodal text extracted from images and videos. OCR-generated text often contains typographical errors, incomplete words, or misread characters. Similarly, ASR-generated transcripts may include inaccuracies caused by background noise, dialect variations, or rapid speech. These factors can negatively impact model performance if not properly addressed.

The results show that the ensemble model is more resilient to noise than the individual classifiers. CNN performed reasonably well with short OCR phrases, while BiLSTM managed longer ASR transcripts more effectively. IndoBERT, however, demonstrated strong robustness in interpreting noisy text due to its pretrained contextual representations. By combining the predictions from all three models, the ensemble was able to compensate for weaknesses in each individual classifier, resulting in higher reliability when processing imperfect multimodal inputs.

3.4. Identification of Implicit Gambling Indicators

A significant portion of the dataset consisted of implicit or disguised advertisements that did not explicitly mention gambling terms. Examples include coded language, stylized abbreviations, or indirect promotional phrases that only experienced users would recognize. IndoBERT showed the strongest performance in identifying such content, likely due to its ability to understand semantic relationships across entire sentences. BiLSTM also contributed through its sequential pattern learning, while CNN showed limitations in this area. The ensemble model captured these implicit cues more effectively than the individual models. This finding highlights the importance of integrating contextual and sequential learning mechanisms, especially in cases where explicit keywords are intentionally avoided to bypass platform moderation systems.

3.5. Discussion

The experimental results underscore the significant advantages of using a multimodal ensemble approach for detecting online gambling advertisements in Indonesian social media. The integration of multiple models—CNN, BiLSTM, IndoBERT, and AutoML—each with distinct strengths, proved to be more effective than relying on any single classifier, particularly when dealing with diverse and noisy content sources.

The comparative performance of individual models reveals the complementary roles each one plays in addressing different aspects of the detection task. CNN demonstrated its strength in identifying explicit gambling-related phrases but struggled with longer, more complex sentences or intricate contextual structures. BiLSTM, on the other hand, excelled in recognizing sequential patterns and dependencies, making it particularly effective in capturing implicit gambling cues embedded within informal or conversational language. IndoBERT stood out as the top performer, thanks to its ability to interpret the broader context and detect subtle, often disguised, gambling-related content—such as coded language, abbreviations, or slang. AutoML, leveraging automated model selection and optimization, performed competitively but ultimately could not match IndoBERT's contextual understanding in nuanced cases.

Combining these models into a multimodal ensemble led to superior overall performance. The ensemble model benefited from the respective strengths of each individual classifier, achieving higher accuracy (95.42%), precision (95.00%), recall (95.62%), and F1-score (95.31%) than any of the standalone models. This confirmed that a hybrid approach—one that blends local pattern recognition, sequential learning, and contextual understanding—offers a more robust and reliable solution, especially in the face of diverse and noisy data typical of social media platforms.

A deeper look at the precision, recall, and F1-score further demonstrates the strengths of the ensemble approach. CNN excelled at precision but was prone to false positives, especially in instances where gambling-related content was embedded in clickbait-style phrases or non-explicit contexts. BiLSTM made strides in recall, capturing more gambling-related instances by focusing on longer contextual dependencies, though it still had limitations in precision. IndoBERT achieved a balanced performance, with high values for both precision and recall, reflecting its superior ability to understand context and subtle gambling cues.

The ensemble model showed the highest F1-score, signaling its balanced handling of false positives and false negatives. The use of weighted soft-voting—where predictions from more accurate models were given greater influence—played a key role in this achievement. By combining CNN's strength in detecting explicit content, BiLSTM's capability to capture sequential relationships, and IndoBERT's deep contextual understanding, the ensemble model effectively mitigated the weaknesses of each individual model, resulting in better classification outcomes.

Social media content often includes noisy text generated by Optical Character Recognition (OCR) or Automatic Speech Recognition (ASR) technologies, which can introduce errors such as typos, misreads, or incomplete words. The ensemble model demonstrated significant robustness to such noise, outperforming the individual classifiers. For instance, while CNN handled short OCR phrases reasonably well and BiLSTM coped with longer ASR transcripts, IndoBERT showed exceptional resilience, utilizing its pretrained contextual embeddings to accurately interpret noisy text. By combining the strengths of each model, the ensemble was able to overcome OCR and ASR-related imperfections, ensuring higher reliability in the final classification.

One of the major challenges in detecting online gambling advertisements is identifying implicit or disguised content, which may avoid explicit gambling-related keywords. IndoBERT emerged as the most effective model for recognizing such subtle indicators, leveraging its capacity to understand the full semantic context of sentences. BiLSTM also contributed through its sequential learning capabilities, but CNN was less effective in identifying such hidden cues. The ensemble model performed best in capturing implicit gambling content, confirming that a combination of contextual and sequential learning strategies is crucial for accurately identifying these hidden advertisements.

The multimodal ensemble approach provides a highly effective and robust solution for detecting online gambling advertisements in Indonesian social media. By leveraging the complementary strengths of CNN, BiLSTM, IndoBERT, and AutoML, the ensemble model achieved superior classification results, especially when handling noisy, multimodal, and implicit content. This approach offers a promising tool for automated content moderation in dynamic and rapidly evolving social media environments.

4. CONCLUSION

This study proposed a multimodal ensemble deep learning framework for detecting Indonesian online gambling advertisements across text, images, and video content. The experimental results show that combining CNN, BiLSTM, and IndoBERT through a weighted soft-voting strategy improves overall performance compared to individual models. The ensemble approach demonstrates higher accuracy, stronger robustness to noisy OCR and ASR outputs, and better detection of implicit gambling indicators. These findings highlight the effectiveness of integrating multiple feature streams and model architectures for handling diverse social media content. Future work may focus on improving multimodal extraction quality, expanding the dataset, and exploring more advanced fusion techniques to further enhance system performance.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Universitas Bina Darma for providing academic support, research facilities, and access to the laboratory environment that enabled the successful completion of this study. The authors also acknowledge the contributions of the annotation team who assisted in labeling the dataset and ensuring the reliability of the ground truth used for model development and evaluation. Their diligent work played a crucial role in the quality of the experimental results. The authors further extend their gratitude to the developers of the open-source tools used in this research, including Word2Vec, IndoBERT, AutoGluon, and the OCR and ASR libraries that facilitated multimodal text extraction. These tools greatly contributed to the advancement of the proposed detection framework. Lastly, the authors appreciate the constructive feedback provided by colleagues and reviewers during the research and manuscript preparation process. Their insights have significantly improved the clarity, methodology, and overall quality of this work.

REFERENCES

[1] Q. Yang, Y. Wang, M. Song, Y. Jiang, and Q. Li, "Sonic strategies: unveiling the impact of sound features in short video ads on enterprise market entry performance," *J. Business-to-bus. Mark.*, vol. 32, no. 1, pp. 95–116, 2025.

[2] A. Shalaby, "Classification for the digital and cognitive AI hazards: urgent call to establish automated safe standard for protecting young human minds," *Digit. Econ. Sustain. Dev.*, vol. 2, no. 1, p. 17, 2024.

[3] J. Singer, A. Wöhr, and S. Otterbach, "Gambling operators' use of advertising strategies on social media and their effects: A systematic review," *Curr. Addict.*

Reports, vol. 11, no. 3, pp. 437–446, 2024.

[4] K. Ataallah *et al.*, "Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens," *arXiv Prepr. arXiv2404.03413*, 2024.

[5] D. Liu *et al.*, "What Is That Talk About? A Video-to-Text Summarization Dataset for Scientific Presentations," *arXiv Prepr. arXiv2502.08279*, 2025.

[6] M. E. Almandouh, M. F. Alrahmawy, M. Eisa, M. Elhoseny, and A. S. Tolba, "Ensemble based high performance deep learning models for fake news detection," *Sci. Rep.*, vol. 14, no. 1, p. 26591, 2024.

[7] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 37979–38007, 2024.

[8] 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, 2024.

[9] A. F. Hidayatullah, "Code-Mixed Sentiment Analysis on Indonesian-Javanese-English Text using Transformer Models," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, 2024, pp. 340–345.

[10] A. NANYONGA, K. F. Joiner, U. Turhan, and G. Wild, "Comparative Analysis of Bert, Cnn, and Lstm Models for Classifying Aviation Safety Incidents in Australia," *Cnn, Lstm Model. Classifying Aviat. Saf. Incidents Aust.*.

[11] S. Albelali and M. Ahmed, "Evaluating the Sensitivity of BiLSTM Forecasting Models to Sequence Length and Input Noise," *arXiv Prepr. arXiv2512.06926*, 2025.

[12] C. Shaw, P. LaCasse, and L. Champagne, "Exploring emotion classification of indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 22, 2025.

[13] S. Lecheheb, S. Boulehouache, and S. Brahimi, "Optimized automated analysis using AutoGluon-driven deep learning for advancing self-adaptive systems," *Computing*, vol. 107, no. 11, pp. 1–24, 2025.

[14] A. Pandey, J. Singh, and M. Kaur, "Bridging Text and Speech for Emotion Understanding: An Explainable Multimodal Transformer Fusion Framework with Unified Audio–Text Attribution," *J. Intell.*, vol. 13, no. 12, p. 159, 2025.