





A Deep Learning and AutoML-Based Multimodal Text Extraction Framework for Detecting Online Gambling Advertisements in Indonesian Social Media

Ari Muzakir^{1,2*}, Usman Ependi², Suyanto¹

¹ Faculty of Science and Technology, Bina Darma University, Palembang 30111, Indonesia

² Master of Informatics Engineering Program, Postgraduate Program, Bina Darma University, Palembang 30111, Indonesia

Corresponding Author Email: arimuzakir@binadarma.ac.id

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ijssse.150908>

ABSTRACT

Received: 18 August 2025

Revised: 20 September 2025

Accepted: 25 September 2025

Available online: 30 September 2025

Keywords:

automatic speech recognition (ASR), AutoML, deep learning, harmful content moderation, Indonesian social media content detection, multimodal text extraction, online gambling detection, optical character recognition (OCR)

The growing presence of online gambling advertisements on social media threatens digital security and complicates content moderation, particularly in multilingual and noisy environments such as Indonesian platforms. This study proposes a deep learning and AutoML-based multimodal text extraction framework that integrates textual posts, optical character recognition (OCR) from images, and automatic speech recognition (ASR) from videos for comprehensive gambling content detection. Four approaches were evaluated: Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Bidirectional Encoder Representations from Transformers (BERT), and AutoML (AutoGluon). Experiments were conducted under standardized preprocessing and hyperparameter-optimized conditions, with performance assessed using accuracy, precision, recall, and F1-score. The RNN achieved the best results (accuracy 93.07%, precision 93.22%, recall 92.87%, F1-score 93.04%), followed by CNN (accuracy 92.80%, F1-score 92.59%) and AutoML (accuracy 90.19%, F1-score 90.18%). BERT underperformed (accuracy 68.12%, F1-score 68.18%) due to limited domain adaptation to noisy Indonesian text. Error analysis revealed three major challenges: implicit promotional language, OCR/ASR transcription noise, and contextual ambiguity. These findings demonstrate the effectiveness of multimodal text integration combined with optimized deep learning models, offering a scalable solution for harmful content moderation and digital security in social media.

1. INTRODUCTION

The rapid growth of social media platforms has fundamentally reshaped how people interact, exchange information, and access entertainment. While this digital transformation brings numerous benefits, it has also facilitated the circulation of illegal and harmful content, including online gambling advertisements. Globally, the online gambling industry reached an estimated market value of USD 63.53 billion in 2022 and is projected to grow further in the coming years. In Indonesia, where internet penetration exceeded 78% in 2023 and social media usage ranks among the highest in Southeast Asia [1], online gambling has become a serious societal and legal concern [2]. Although prohibited under national law, operators exploit the anonymity and vast reach of platforms such as Facebook, X (formerly Twitter), Instagram, and YouTube to promote gambling services to a wide audience, including vulnerable users such as teenagers.

The risks associated with online gambling advertisements extend beyond legal violations. They include social harm, such as fostering addictive behaviors, and economic harm, such as financial losses among players [3]. Moreover, gambling-related crimes and fraudulent schemes often follow, making detection and prevention a matter of public interest. Law enforcement agencies and platform moderators face

substantial challenges in identifying such content [4], especially because promotional messages are often implicit, using coded words, local slang, or persuasive narratives to evade automated detection systems [5].

Detecting gambling-related content in text-based social media is inherently challenging due to the high volume, linguistic diversity, and noise in online posts. Messages often contain informal language [6], abbreviations, emojis, and code-switching between Indonesian and English. The morphological complexity of the Indonesian language, with its extensive affixation system, further complicates accurate classification. In addition, gambling promotions are not always embedded in plain text. They can appear in images, embedded within videos, or hidden in spoken audio. This limitation in prior research has left a detection gap, as most automated systems only process raw text fields and ignore content from other modalities.

Several approaches have been proposed to address harmful content detection, yet each carries notable limitations. Traditional machine learning methods, such as TF-IDF combined with Random Forest, have shown strong precision for gambling-related text in Indonesian Twitter data [2]. However, these models rely heavily on handcrafted features and often fail to generalize across diverse platforms. Convolutional Neural Networks (CNNs) are effective at

capturing local n-gram patterns in text, yet they are limited in modeling long-range dependencies [7, 8]. Recurrent Neural Networks (RNN), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), can better capture sequential dependencies and context but are slower to train and sensitive to noise in informal text [9, 10]. Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and its Indonesian variants (e.g., IndoBERT) offer state-of-the-art contextual representation [11], but they often underperform when applied to noisy, slang-heavy, and code-switched Indonesian data without extensive domain-specific pretraining. AutoML frameworks like AutoGluon have been increasingly adopted for text classification tasks due to their automated model selection and hyperparameter tuning capabilities [12]. However, their default pipelines may fail to fully capture linguistic nuances in morphologically rich and informal languages such as Indonesian. Prior multimodal studies also tend to focus on isolated modalities, with optical character recognition (OCR) struggling in low-resolution or stylized images [13], and automatic speech recognition (ASR) models suffering from transcription errors in noisy or dialect-rich audio [14].

Based on this review, two major research gaps are identified. The first is the absence of a comprehensive multimodal-to-text framework for Indonesian social media that integrates native text, OCR-extracted text from images, and ASR-transcribed speech into a unified representation. The second is the lack of systematic cross-model evaluations of CNN, RNN, Transformer-based models, and AutoML conducted under consistent preprocessing and hyperparameter settings.

To address these gaps, this study proposes a unified multimodal-to-text extraction framework that consolidates gambling-related content from native text, images, video frames, and audio transcripts into a standardized textual representation. On top of this dataset, a comparative cross-model analysis is conducted to evaluate CNN, LSTM, BiGRU, BERT, and AutoML (AutoGluon) under identical conditions, enabling a fair assessment of their strengths, weaknesses, and trade-offs in terms of accuracy, efficiency, and interpretability.

The contributions of this study are as follows: (1) it introduces a unified multimodal-to-text pipeline that integrates OCR and speech-to-text for detecting gambling-related content across text, images, video frames, and audio; (2) develops a domain-specific Indonesian dataset of 15,286 annotated samples from multiple platforms; (3) conducts a systematic comparison of CNN, LSTM, BiGRU, BERT, and AutoML under consistent settings; and (4) provides practical insights for real-world deployment that balance detection performance with operational efficiency.

2. RELATED WORKS

2.1 Multimodal text extraction in content moderation

Illicit content detection in social media has historically relied on native text analysis. However, gambling advertisements frequently appear in images, videos, or audio streams, making unimodal approaches insufficient. To address this, OCR has been applied to extract embedded text from images and video frames [15]. For instance, Thapa et al. [16] demonstrated that detecting hate speech in memes requires multimodal strategies combining visual and textual features.

OCR effectively reveals hidden textual signals invisible to traditional text classifiers, though its performance often degrades with low-resolution inputs, stylized fonts, or cluttered backgrounds [13].

In video content, extracting text from frames provides additional cues for moderation. Xu et al. [17] reported that traditional video-based fake news detection often relies on unimodal features, while more recent video-OCR methods highlight the importance of analyzing embedded text to improve multimodal understanding. Nonetheless, these methods face challenges such as motion blur and high computational costs.

For audio, moderation systems frequently employ ASR to convert speech into text prior to classification. Bell et al. [18] showed that toxic speech detection pipelines typically depend on ASR, but accuracy is highly sensitive to noise, accents, and recording variability. In the Indonesian context, Adila et al. [19] highlighted the scarcity of diverse ASR datasets, while code-switching between Indonesian and English continues to degrade transcription quality [20].

Overall, most studies focus on English and unimodal settings, leaving a gap for Indonesian-language research. To date, limited work has attempted to integrate OCR, video frame text extraction, and ASR into a unified multimodal pipeline for harmful content detection [21].

2.2 Comparative cross-model analysis in text classification

Online gambling detection in Indonesian social media is still in its early stages. Prior work notes that informal grammar, slang, code-mixing, and limited annotated data make classification especially challenging [2]. Most studies employ individual models, without systematic comparisons across model families.

Early methods relied on handcrafted features such as bag-of-words and TF-IDF with classifiers like SVM or Random Forest [22]. For example, Perdana et al. [2] reported $\approx 96\%$ precision for gambling promotion detection on Twitter using TF-IDF + Random Forest. While efficient, such methods require extensive feature design and lack adaptability to nuanced or evolving language.

Deep learning has reduced the need for manual feature engineering. CNNs are effective at detecting local n-gram patterns, slang, and emoticons [8], but they are inherently limited in modeling long-range dependencies. RNN-based models such as LSTM and GRU better capture sequential context [10], yet they train more slowly and are sensitive to noisy or lengthy inputs. Transformer-based models like BERT have become the standard in NLP, learning bidirectional context and achieving strong results in Indonesian variants such as IndoBERT [11]. However, these models demand substantial computational resources and offer limited interpretability, which poses challenges for real-time moderation.

AutoML frameworks, including AutoGluon and H2O AutoML, have recently gained traction for their ability to automate feature engineering, model selection, and hyperparameter optimization. In text classification tasks, AutoML can match the performance of manually tuned deep learning models while reducing development effort [12]. Still, their default pipelines struggle with morphologically rich and informal Indonesian text, reinforcing the need for domain-specific adaptation [23].

2.3 Research gap

Based on the reviewed literature, two critical research gaps can be identified. The first is the absence of a comprehensive multimodal-to-text extraction pipeline for Indonesian social media that integrates OCR, ASR, and native text into a unified workflow. The second is the lack of systematic comparative evaluations across CNN, RNN, Transformer-based models, and AutoML conducted under consistent experimental conditions.

This study addresses these gaps by proposing a fully integrated multimodal-to-text pipeline and conducting an extensive cross-model benchmarking analysis. By aligning the experimental setup across all models, the study ensures a fairer comparison and provides actionable insights for both academia and industry. A consolidated overview of representative prior studies, their methodological focus, and key findings is presented in Table 1. The table highlights existing contributions while emphasizing persisting limitations that motivate the need for this study.

As shown in Table 1, prior studies demonstrate that traditional machine learning methods such as TF-IDF

combined with Random Forest are efficient and can achieve high precision, but they depend heavily on handcrafted features and often fail to generalize to informal or evolving language patterns. Deep learning models alleviate this limitation by automatically learning feature representations: CNN capture local n-grams effectively but struggle with long-range dependencies, RNN model sequential context but are computationally slower and sensitive to noise, while Transformers such as BERT achieve state-of-the-art results but require substantial resources and lack interpretability. AutoML frameworks like AutoGluon further reduce development effort by automating model selection and hyperparameter tuning, yet their default pipelines may not fully capture the linguistic complexity of Indonesian, making domain-specific adaptation essential. These trade-offs highlight that no single approach is universally optimal. This study addresses these gaps by proposing a unified multimodal-to-text pipeline and conducting a systematic cross-model evaluation of CNN, RNN, Transformer-based, and AutoML approaches under consistent experimental settings, thereby providing both theoretical and practical insights for Indonesian online gambling content detection.

Table 1. Summary of relevant prior studies in multimodal content detection and cross-model text classification

Study	Domain	Data Type	Methods	Key Advantages	Limitations
Perdana et al. [2]	Indonesian gambling detection	Text (Twitter)	TF-IDF + Random Forest	High precision ($\approx 96\%$) on gambling ads	Limited generalization, needs manual feature design
Thapa et al. [16]	Hate speech in memes	Images	OCR + multimodal features	Captures hidden embedded text in images	Degrades with low-resolution or stylized fonts
Xu et al. [17]	Video fake news	Video frames	Frame sampling + OCR	Extends detection to multimodal video content	Sensitive to motion blur, computationally costly
Bell et al. [18]	Toxic speech	Audio + text	ASR transcription + classification	Enables spoken content moderation	Transcription errors from noise, accents, and variability
Adila et al. [19]	Indonesian ASR	Audio (speech)	End-to-end ASR models	Builds Indonesian resources for speech-to-text	Scarcity of diverse datasets, poor code-switching handling
Wilie et al. [24]	Indonesian NLP	Text	IndoBERT	Strong adaptation for Bahasa Indonesia	Large size, slower inference
Anitha et al. [25]	NLP tasks	Text	AutoGluon (AutoML)	Automated model/hyperparameter selection	Domain-specific tuning still required

3. METHODOLOGY

This study follows a structured research pipeline consisting of five main stages: (1) data collection and annotation, (2) preprocessing, (3) feature representation, (4) model development, and (5) evaluation.

The overall workflow is shown in Figure 1. The proposed methodology incorporates a multimodal-to-text extraction process to unify content from different data sources into a single text-based format before classification.

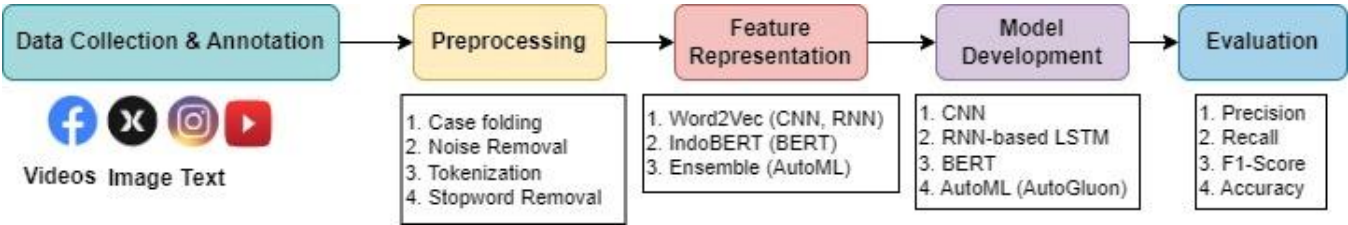


Figure 1. Research workflow for online gambling advertisement detection in social media

3.1 Data collection and annotation

Data were collected from four major social media platforms: Facebook, X (formerly Twitter), Instagram, and YouTube, following the sequential workflow illustrated in Figure 2. The process began with raw data acquisition, ensuring diversity

across text posts, images, and videos. In the second stage, video content was processed by extracting keyframes at fixed intervals, after which OCR was applied to retrieve any embedded textual elements. Similarly, static images, including promotional banners and graphics, were processed through OCR to extract written content. In the third stage, the audio

streams from videos were transcribed into text using an ASR system.

This multimodal extraction pipeline enabled the inclusion of gambling-related promotions that often appear in non-textual formats, a common evasion tactic employed by online gambling operators. The final stage integrated all outputs into a unified text corpus, consisting of three primary modalities: (1) native text posts directly obtained from captions, comments, and descriptions, (2) textual content extracted from images and video frames through OCR, and (3) speech-to-text transcriptions from video audio. This standardized representation served as the foundation for subsequent preprocessing and model training. To ensure completeness, video audio streams were also processed independently through ASR before being merged with OCR and native text, guaranteeing that no modality was overlooked.

After the collection and extraction process, the resulting corpus comprised 15,286 textual entries. Each entry was manually annotated by eight trained volunteers who had prior experience or exposure to online gambling activities. This background was considered beneficial for recognizing implicit gambling-related cues.

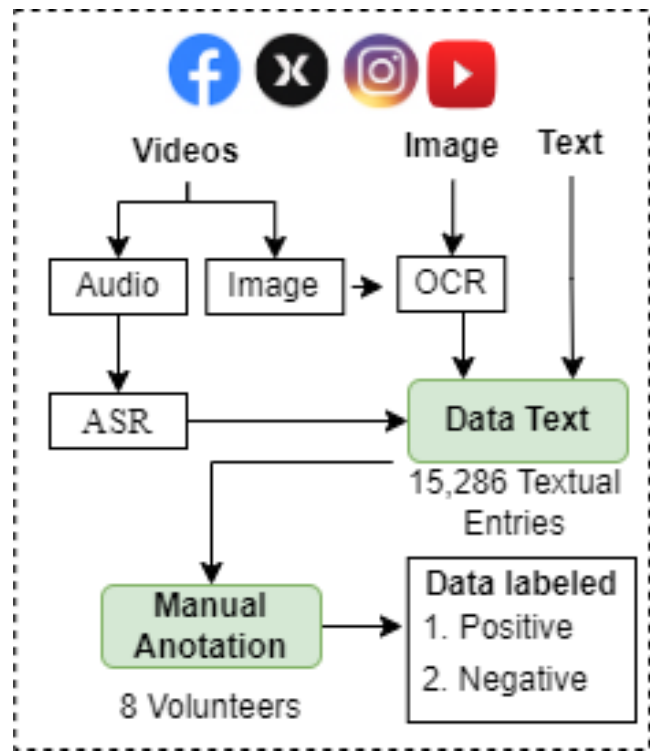


Figure 2. Data extraction workflow from multimodal social media content into unified text format

Table 2. Presents examples of annotated data for both classes

Example Text	Label	Remarks
"Let's try to guess your lucky number today"	Positive	Implicit gambling intent
"Big discounts on all fashion products"	Negative	Non-gambling promotional content
"You can earn millions of rupiah just from home"	Positive	Gambling implication is subtle

The annotators labeled the dataset into two categories: Positive, referring to text explicitly or implicitly promoting or discussing online gambling activities, and Negative, referring to text unrelated to gambling, including general conversation,

unrelated promotions, or neutral content. Illustrative examples of each category are presented in Table 2, providing clarity on the annotation guidelines applied during dataset construction.

The distribution pattern is illustrated in Figure 3, which provides a visual overview of class balance across the dataset. The final label distribution after quality control was as follows: Positive contained 8,499 entries (55.6%), and Negative contained 6,787 entries (44.4%). This distribution indicates a slight dominance of positive samples, with around 11.2% more than negative samples, which is not considered severely imbalanced.

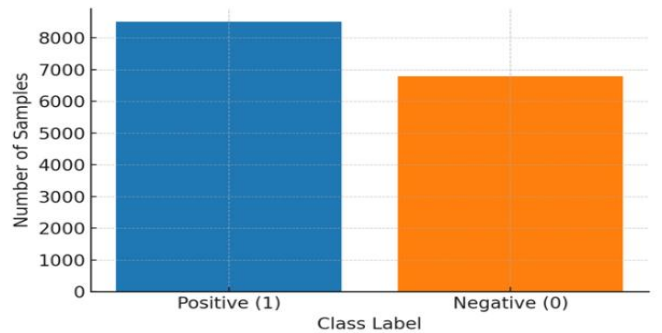


Figure 3. Label distribution in the dataset

The dataset was stratified and split into three subsets, consisting of 75 percent for training, 10 percent (from the training set) for validation, and 15 percent for testing, ensuring that the class distribution was preserved in all subsets.

3.2 Preprocessing

The preprocessing stage aimed to clean and normalize the collected textual data to ensure consistency and improve model performance. First, case folding was applied by converting all characters to lowercase, eliminating variations caused by capitalization and standardizing word forms. Second, noise removal was performed to remove non-alphanumeric characters, such as emojis, symbols, and URLs. This step also involved cleaning repetitive punctuation marks and excessive whitespace. Third, tokenization was carried out to split sentences into individual tokens based on whitespace, with additional adjustments for separating punctuation from words. Finally, stopwords removal was conducted using the Sastrawi stopwords list for Indonesian, removing frequently occurring but semantically uninformative words such as *dan*, *yang*, and *atau*. These operations reduced data sparsity and enhanced the signal-to-noise ratio, allowing models to focus on meaningful terms.

3.3 Feature representation

Two embedding strategies were employed: Word2Vec and IndoBERT.

- CNN and RNN (LSTM): Word2Vec embeddings (200-dimensional vectors pretrained on large Indonesian corpora) preserved semantic relationships, generating sentence-level sequences fed into neural architectures.
- BERT-based model: IndoBERT-base-p1 produced contextual embeddings through WordPiece tokenization, with sequences padded or truncated to length 128. Unlike static embeddings, IndoBERT generated dynamic context-sensitive vectors.

- AutoML (AutoGluon): handled feature representation automatically with built-in preprocessing (tokenization, BoW, TF-IDF, n-grams) and optionally pretrained embeddings in ensembles.

This multi-representation strategy ensured fairness by optimizing features for each model type, enabling a consistent basis for comparison.

3.4 Model development

This study implemented and compared four distinct modeling approaches for classifying online gambling advertisements in Indonesian-language social media content. Two approaches were manually designed deep learning models (CNN and RNN with LSTM), one was a transformer-based model (IndoBERT), and one was an automated machine learning approach (AutoGluon). All models were trained and evaluated using the same dataset and performance metrics to ensure a fair comparison. All models were trained and evaluated using the same dataset and performance metrics to ensure a fair comparison, as shown in Table 3.

Table 3. Architectures and parameters of classifiers

Model	Architecture and Parameters
CNN-Based Classifier	<ol style="list-style-type: none"> 1. Embedding layer: pretrained Word2Vec, 200-dim vectors 2. Conv1D: 128 filters, kernel size = 5, activation = ReLU 3. Max-Pooling layer 4. Dense layer: fully connected, dropout = 0.5 5. Output: sigmoid layer 6. Loss: Binary Cross-Entropy 7. Optimizer: Adam, LR = 0.001 8. Training: 10 epochs, batch size = 64, early stopping
RNN-Based Classifier (LSTM)	<ol style="list-style-type: none"> 1. Embedding layer: pretrained Word2Vec, 200-dim vectors 2. LSTM layer: 128 hidden units 3. Dense layer: 64 units, activation = ReLU 4. Dropout = 0.5 5. Output: sigmoid layer 6. Loss: Binary Cross-Entropy 7. Optimizer: Adam, LR = 0.001 8. Training: 10 epochs, batch size = 64, early stopping
Transformer-Based Classifier (IndoBERT)	<ol style="list-style-type: none"> 1. Tokenizer: WordPiece, max length = 128 2. Encoder: <i>indobert-base-pl</i> 3. Dense layer: dropout = 0.3 (pooled output) 4. Output: sigmoid layer 5. Loss: Binary Cross-Entropy 6. Optimizer: AdamW, LR = 2e-5 7. Training: 4 epochs, batch size = 32
AutoML (AutoGluon)	<ol style="list-style-type: none"> 1. Framework: AutoGluon 2. Automated preprocessing: tokenization, feature extraction, missing value handling 3. Candidate models: GBM, Random Forest, Bagging Ensembles, Feed-forward NN 4. Stacked ensemble generation 5. Training time: max 300 sec/trial

3.4.1 CNN-based classifier

The CNN architecture was designed for binary text classification, using sequences of pretrained Word2Vec embeddings as input. Each token was mapped to a 200-dimensional dense vector. The architecture consisted of:

- An embedding layer for Word2Vec vector mapping
- A 1D convolutional layer with 128 filters and a kernel

size of 5, followed by ReLU activation

- A max-pooling layer to reduce dimensionality and capture local patterns
- A fully connected dense layer with dropout regularization (dropout rate of 0.5)
- A final sigmoid output layer for binary classification

The model was trained using the binary cross-entropy loss function and optimized with Adam at a learning rate of 0.001. Training was conducted for 10 epochs with a batch size of 64, and early stopping was applied based on validation loss to prevent overfitting. The CNN was implemented using *TensorFlow/Keras*.

3.4.2 RNN-based classifier (LSTM)

The RNN model was implemented using a LSTM architecture, also taking sequences of 200-dimensional Word2Vec embeddings as input. The architecture included:

- An embedding layer for Word2Vec vector mapping
- An LSTM layer with 128 hidden units to capture sequential dependencies
- A dense layer with 64 units and ReLU activation
- Dropout regularization (dropout rate of 0.5) to mitigate overfitting
- A sigmoid output layer for binary classification

The model was trained with binary cross-entropy loss and the Adam optimizer (learning rate 0.001) for 10 epochs, batch size 64, with early stopping on validation loss. This architecture was chosen for its ability to capture long-range dependencies in sequential data.

3.4.3 BERT-based classifier (IndoBERT)

For transformer-based classification, the pretrained *indobert-base-pl* model was fine-tuned on the gambling detection dataset. Input sentences were tokenized using the WordPiece tokenizer, padded or truncated to a maximum length of 128 tokens. The classification head consisted of:

- The IndoBERT encoder
- A dense layer with dropout (0.3) applied to the pooled output
- A final dense layer with a sigmoid activation for binary classification

The model was fine-tuned using binary cross-entropy loss and AdamW optimizer with a learning rate of 2e-5 for 4 epochs, batch size 32. The Hugging Face Transformers library was used for implementation.

3.4.4 AutoML (AutoGluon)

The AutoML approach utilized the AutoGluon framework, which provided automated preprocessing, model selection, and hyperparameter tuning. AutoGluon was configured with a maximum training time of 300 seconds per trial. Candidate models included Gradient Boosting Machines (GBM), Random Forests, Bagging Ensembles, and feed-forward neural networks. AutoGluon handled tokenization, feature extraction, and missing value imputation automatically, as well as generating stacked ensembles to improve performance.

By employing CNN, LSTM-based RNN, IndoBERT, and AutoML, this study enabled a direct comparison between manual deep learning architectures, transformer-based language models, and automated machine learning systems. All models were trained on the same preprocessed dataset and evaluated using accuracy, precision, recall, and F1-score to ensure consistency in performance assessment.

3.4.5 Evaluation

To evaluate the effectiveness of the proposed models in detecting online gambling advertisements, this study employed a set of standard classification metrics widely used in binary text classification tasks: accuracy, precision, recall, and F1-score. These metrics collectively provide a comprehensive performance assessment from different perspectives, capturing both correctness and robustness in classification. The evaluation metrics are calculated using Eqs. (1)-(4):

$$F1\text{-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

$$\text{Accuracy} = \frac{TP + FN}{TP + FN + TN + FP} \quad (2)$$

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

where, TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

In addition to the numerical evaluation, a confusion matrix was generated for each model to visualize classification performance and detect potential misclassification patterns between the positive and negative classes. This visual analysis is particularly important in the gambling advertisement detection task, as certain posts may contain implicit or coded promotional language that is harder to identify. Furthermore, error analysis was performed by manually reviewing a subset of misclassified instances, enabling the identification of recurring linguistic structures or contextual cues that contributed to incorrect predictions.

3.4.6 Hyperparameter tuning

To ensure transparency and reproducibility, a systematic hyperparameter tuning process was conducted for all models. A grid search approach was applied on the training and validation sets, and the optimal configurations were selected based on the highest validation F1-score while maintaining a balanced precision-recall trade-off.

In the case of CNN and RNN models, different batch sizes (32, 64, and 128), learning rates (0.001, 0.0005, and 0.0001), and training epochs (5, 10, and 15) were explored. The configuration with a batch size of 64, learning rate of 0.001, and 10 epochs yielded the most favorable balance between convergence speed and generalization.

For the BERT-based model, experiments were conducted with batch sizes of 16 and 32 combined with learning rates of $5e-5$, $3e-5$, and $2e-5$. The best results were obtained using a batch size of 32 and a learning rate of $5e-5$ for 5 epochs. This setup was chosen to address GPU memory constraints while minimizing the risk of overfitting.

With AutoML (AutoGluon), the maximum training time per trial was fixed at 300 seconds. This allowed adequate exploration of candidate models while keeping computational costs practical.

Across all models, a total of 54 hyperparameter combinations were evaluated. This systematic procedure and the rationale behind parameter selection ensured that the comparative evaluation was fair, transparent, and reproducible.

4. RESULTS AND DISCUSSION

4.1 Experimental setup and hyperparameter tuning

In this study, a series of controlled experiments were conducted to evaluate the performance of the proposed unified multimodal-to-text framework for detecting online gambling advertisements in Indonesian social media. The primary objective was to determine the optimal configuration of deep learning and AutoML-based models, ensuring robust classification performance across various linguistic and contextual challenges. Two experimental scenarios were designed:

1. Model architecture evaluation. In which the performance of CNN, LSTM, BERT, and AutoML models was compared under standardized preprocessing and dataset conditions.
2. Hyperparameter optimization. Aimed at identifying the best set of hyperparameters for each model to achieve maximum classification accuracy and F1-score.

A standardized preprocessing pipeline was applied to all models, including tokenization, case-folding, stopword removal, and handling of emojis, slang, and code-switching. This ensured that any performance differences were attributable to model capabilities rather than preprocessing variations. The following hyperparameters were explored:

- Batch Size: 32, 64, and 128 were tested. Larger batch sizes improved training speed but sometimes reduced generalization. Batch size 64 was optimal for CNN and LSTM, while BERT used batch size 32 due to memory constraints.
- Learning Rate: CNN and LSTM were tested with 0.001, 0.0005, and 0.0001, while BERT used $5e-5$, $3e-5$, and $2e-5$. The optimal values were 0.001 for CNN/LSTM and $5e-5$ for BERT.
- Epochs: 5, 10, and 15 were tested. CNN and LSTM performed best at 10 epochs, while BERT reached optimal performance at 5 epochs.
- Embedding Representation: CNN and LSTM used 200-dimensional Word2Vec embeddings trained on Indonesian social media corpora. BERT was initialized with *IndoBERT-base-pl* and fine-tuned end-to-end. AutoML used AutoGluon *TabularPrediction* with automatic ensembling.

In Scenario 1, all models were trained using identical datasets and preprocessing to ensure fairness in comparison. In Scenario 2, the best-performing hyperparameter configuration for each model was selected based on the highest validation F1-score and balanced precision-recall.

Preliminary experiments demonstrated that certain hyperparameter configurations consistently produced the highest performance across evaluation metrics. For the CNN and LSTM models, the optimal setup was a batch size of 64, a learning rate of 0.001, and 10 training epochs, providing a good balance between convergence speed and generalization capability. The BERT model achieved its best results with a smaller batch size of 32, a learning rate of $5e-5$, and 5 training epochs, which helped prevent overfitting while effectively leveraging its large pre-trained parameter space. For the AutoML framework, the most effective configuration was obtained by setting the maximum training time per trial to 300 seconds, enabling thorough model exploration and ensembling within practical computational limits.

Across both scenarios, a total of 54 unique model-

hyperparameter combinations were tested, generating more than 200 individual training runs when accounting for cross-validation. The optimal configuration for each model was then carried forward to subsequent evaluation stages, including confusion matrix analysis, class-wise precision-recall computation, and error analysis.

4.2 Model accuracy comparison and learning curve analysis

To evaluate the relative performance of the proposed models, a comparative analysis was conducted on the classification accuracy achieved by RNN, CNN, BERT, and AutoML under their respective optimal hyperparameter configurations. This comparison enables a direct assessment of each model’s ability to address the linguistic challenges of Indonesian social media text, including informal expressions, slang, code-switching, and implicit references to gambling activities.

As illustrated in Figure 4, the RNN model achieved the highest overall accuracy at 93.07%, followed closely by CNN at 92.80% and AutoML at 90.19%, whereas BERT obtained a considerably lower accuracy of 68.12%.

The superior performance of the RNN can be attributed to its capacity to capture long-range sequential dependencies in text, which is particularly advantageous for identifying contextual patterns in gambling-related advertisements. CNN also performed competitively, leveraging its strength in extracting salient n-gram features from text, although it was

slightly less effective than RNN in modeling longer contextual dependencies. AutoML demonstrated robust performance without manual architecture tuning, benefiting from automated feature selection and ensembling strategies. In contrast, BERT’s lower accuracy suggests that domain-specific adaptation is essential for maximizing the potential of pre-trained contextual embeddings in this task.

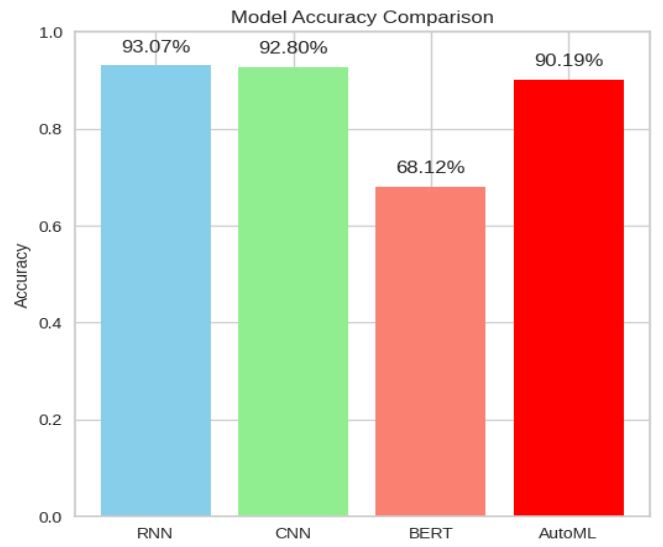


Figure 4. Model accuracy comparison

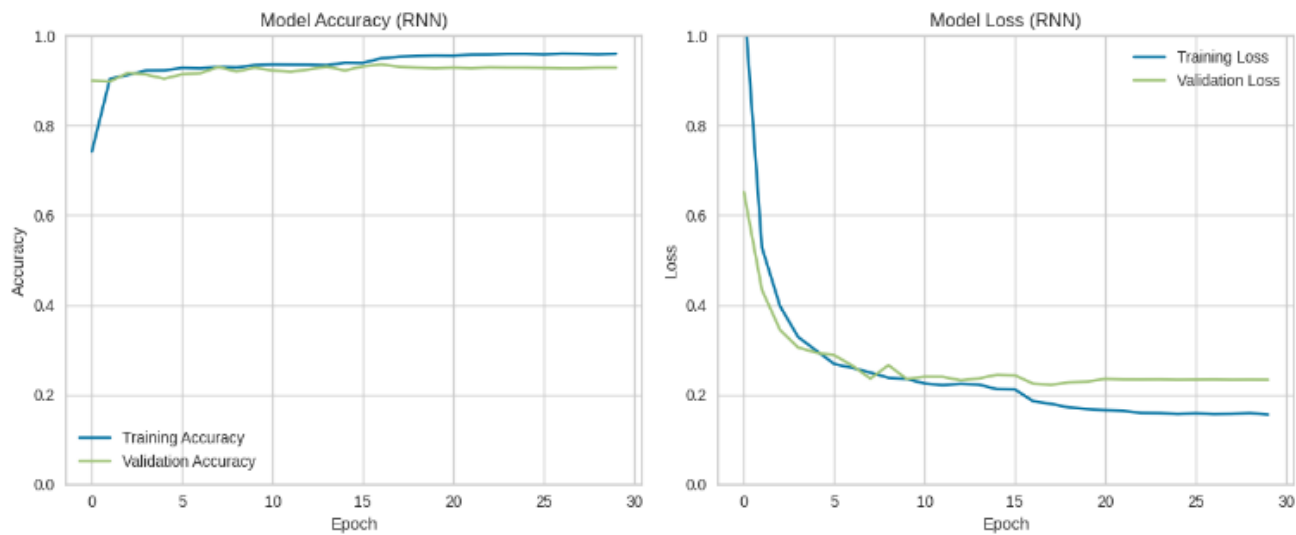


Figure 5. Accuracy and loss curves for the best model

Analysis of Figure 5 reveals that RNN demonstrated the most stable and consistent convergence among all models, reaching peak performance with a final training accuracy of 93.85% and validation accuracy of 93.07%, while maintaining a low final validation loss of 0.172. The minimal gap of only 0.78% between training and validation accuracy indicates strong generalization to unseen data and resilience against overfitting. CNN converged rapidly, reaching its highest validation accuracy of 92.80% by epoch 8 with a validation loss of 0.185; however, minor fluctuations of $\pm 0.4\%$ in validation accuracy after epoch 8 suggested mild overfitting tendencies in later stages. BERT achieved stable convergence with a final validation accuracy of 68.12% and validation loss of 0.423, maintaining a small training–validation accuracy gap

(~1.1%). Nonetheless, the relatively low overall accuracy indicates that while BERT preserved consistency during training, its architecture may require extensive domain-specific fine-tuning to capture the nuanced linguistic patterns of Indonesian social media text. Overall, these results indicate that while CNN and BERT offer competitive and stable training behavior, RNN stands out as the most effective and generalizable model in this domain, combining high accuracy with minimal overfitting risk.

4.3 Error analysis

An in-depth error analysis was conducted to better understand the strengths and weaknesses of each model in

detecting online gambling advertisements in Indonesian social media. This section integrates quantitative performance metrics with qualitative error inspection to provide a comprehensive perspective on model behavior.

The first step of analysis involved evaluating per-class performance metrics. As summarized in Table 4, the RNN model achieved the highest and most balanced results, with both Positive and Negative classes exceeding 93% across precision, recall, and F1-score. The CNN model followed closely, delivering strong results but showing slightly less stability in recall for the Negative class. AutoML produced a relatively balanced performance across classes, while BERT underperformed substantially, with F1-scores of only ~68% for both classes.

Table 4. Per-class performance metrics for all models

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
RNN	Positive	93.22	92.87	93.04
	Negative	92.92	93.27	93.09
CNN	Positive	92.53	92.64	92.59
	Negative	93.07	92.96	93.01
BERT	Positive	68.45	67.92	68.18
	Negative	67.82	68.35	68.08
AutoML	Positive	90.04	90.32	90.18
	Negative	90.33	90.05	90.19

To gain deeper insights, confusion matrix analysis was

carried out. As shown in Figure 6, the RNN model achieved strong performance with only 88 false positives and 71 false negatives, indicating a balanced predictive ability. The CNN model yielded comparable results, with 105 false positives but the lowest number of false negatives (60), highlighting its slightly higher sensitivity to gambling-related content. AutoML also performed competitively, but its false negatives (121) were higher than those of RNN and CNN, suggesting a tendency to overlook positive cases. In contrast, BERT performed the weakest, with 328 false positives and 403 false negatives, reflecting significant difficulty in distinguishing gambling from non-gambling texts. These results confirm that RNN and CNN consistently deliver robust classification, AutoML requires improvements in handling positive cases, and BERT struggles without further domain-specific fine-tuning.

Beyond the numbers, each model showed distinct tendencies. RNN excelled in sequential context modeling, which helped capture implicit gambling patterns, though it occasionally missed very short posts or content heavily tied to visual cues. CNN was particularly effective at detecting short, keyword-rich gambling posts, but sometimes produced false positives in ambiguous contexts due to its reliance on local n-gram features. BERT handled complex sentence structures relatively well, yet struggled with slang and creative variations, leading to frequent misclassifications. AutoML delivered balanced predictions through ensemble learning, though its errors often mirrored those of CNN, especially with contextually ambiguous inputs.

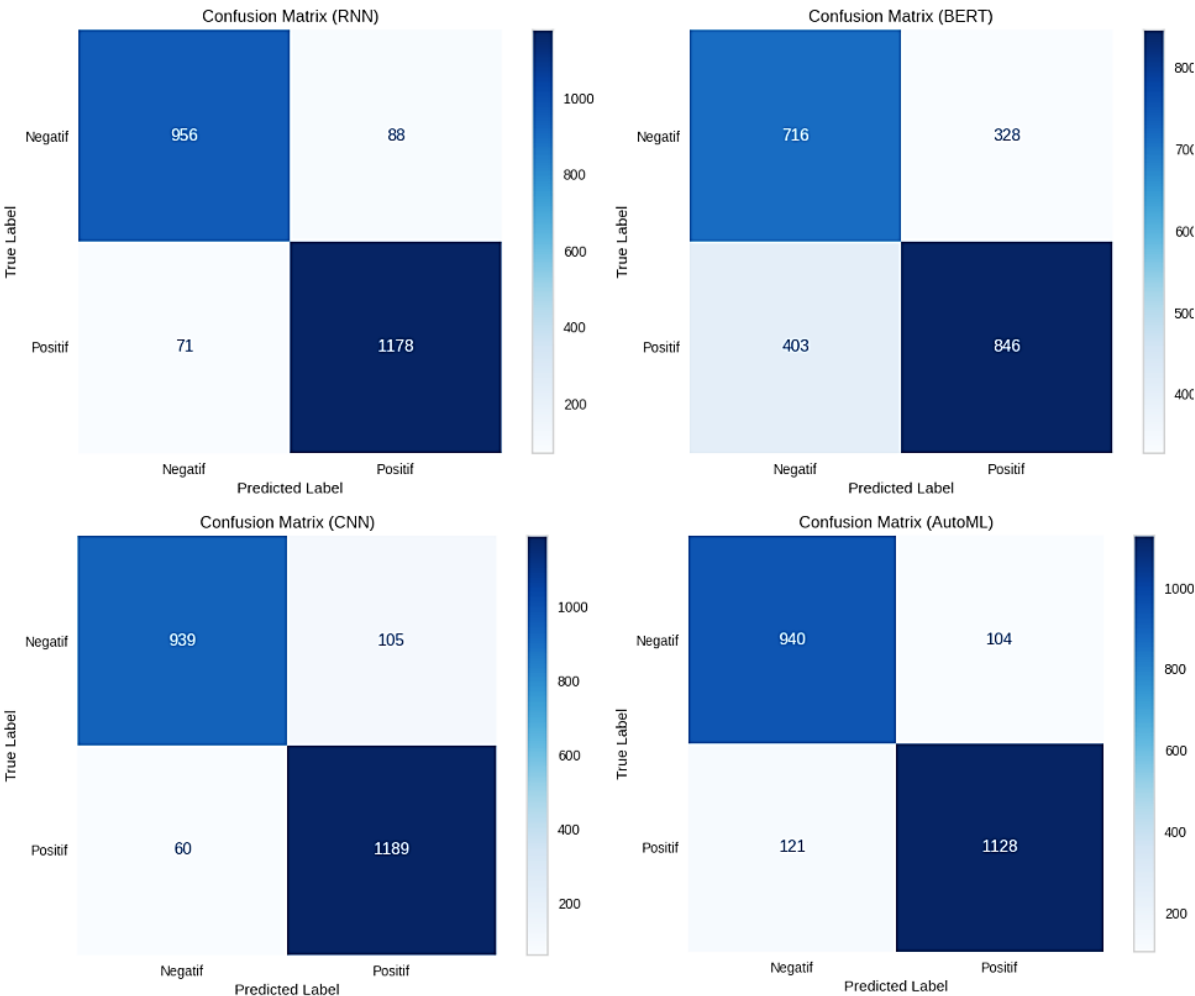


Figure 6. Confusion matrices of RNN, CNN, BERT, and AutoML models on testing data

To further understand the nature of errors, representative misclassified examples were analyzed. As summarized in Table 5, three primary sources of misclassification were identified:

1. Implicit or coded promotional language. Many gambling advertisements used local slang, abbreviations, or code words (e.g., “*jp*”, “*maxwin*”, “*scatter*”) that may not explicitly indicate gambling but are widely understood by target audiences. Models without strong contextual representation tended to misclassify these instances as non-gambling.
2. OCR and speech-to-text noise. Text extracted from images or transcribed from video speech often contained spelling errors, missing words, or extraneous characters. This noise degraded semantic clarity, leading to misclassification even for otherwise explicit gambling promotions.
3. Contextual ambiguity. Some posts contained gambling-related terms used metaphorically or in unrelated contexts (e.g., “*play slot in my heart as a joke*”), causing models to incorrectly label them as positive.

Taken together, these findings highlight two critical

weaknesses shared across models: reliance on explicit keywords and vulnerability to noisy text. Even advanced architectures like BERT struggled with implicit references when contextual evidence was insufficient, while OCR/ASR noise further degraded semantic consistency. Future work could address these issues through domain-specific vocabulary expansion, context-aware phrase mining, and noise-robust pretraining techniques, ultimately improving robustness in real-world multimodal environments.

The observed performance differences can be attributed to the inherent modeling strengths of each approach. CNN primarily captures local n-gram features, which explains its effectiveness in detecting keyword-rich gambling posts, but also its tendency to misclassify metaphorical or ambiguous terms. RNN, by contrast, leverages sequential dependencies, making it more robust to informal expressions and implicit cues distributed across longer text sequences. AutoML, through its ensemble strategies, achieved relatively balanced results, though its reliance on generic feature engineering limited its ability to capture domain-specific linguistic nuances.

Table 5. Representative examples of misclassified instances

Example Text	Actual Label	Predicted Label	Possible Cause of Misclassification
“Profit every day scatter 3x auto withdraw boss”	Positive	Negative	Gambling slang (“scatter”, “withdraw”) not recognized due to limited exposure in training set.
“This slot really makes my heart happy, like finding a soulmate”	Positive	Negative	Metaphorical use of the term “slot” reduced semantic clarity.
“Maxwin only comes from hard work, not from the gambling table”	Negative	Positive	Presence of “maxwin” triggered false positive despite an anti-gambling context.
“Register now via the blue link to get instant maxwin”	Positive	Negative	OCR noise removed a keyword, weakening the gambling-related signal.
“If you lose in the arena, don’t give up... try again!”	Negative	Positive	Competitive gaming terminology overlapped with gambling vocabulary.

Table 6. Comparison with relevant prior studies

Study	Domain	Language	Model	Dataset Size	Accuracy /F1-Score	Key Strengths	Limitations
Perdana et al. [2]	Gambling ads (Twitter)	Indonesian	TF-IDF + Random Forest	~3K tweets	Precision ≈ 96%	Strong gambling detection with handcrafted features	Poor generalization beyond Twitter
Thapa et al. [16]	Hate speech in memes	English	OCR + multimodal features	~10K meme images	F1 ≈ 82%	Captures hidden embedded text in images	Degrades with low-res or stylized fonts
Xu et al. [17]	Video fake news	English	Frame sampling + OCR	~5K video clips	Acc ≈ 85%	Extends detection to multimodal video content	Sensitive to motion blur, computationally costly
Bell et al. [18]	Toxic speech (audio/text)	English	ASR + classification (LSTM/CNN)	~8K audio clips	F1 ≈ 80%	Enables spoken content moderation	Transcription errors due to noise, accents
Adila et al. [19]	Indonesian ASR	Indonesian	End-to-end ASR (DeepSpeech)	~200 hrs speech	WER ≈ 18%	Builds Indonesian ASR resources	Limited dataset diversity, weak in code-switching
Wilie et al. [24]	Indonesian NLP	Indonesian	IndoBERT	~20 + NLP tasks datasets	Acc > 90%	Strong adaptation to Bahasa Indonesia	Large model size, slower inference
Anitha et al. [25]	NLP tasks	English	AutoGluon (AutoML)	~5 NLP datasets	Acc ≈ 88%	Automated feature/model/hyperparameter selection	Still requires domain-specific tuning
This Study	Gambling ads (multimodal)	Indonesian	CNN, RNN, IndoBERT, AutoML	15K (text + image + video)	Acc = 93.07% (RNN), F1 > 92%	Robust multimodal detection; strong balance across classes	Some errors in short/noisy or highly encrypted content

BERT’s underperformance is particularly notable. Although transformer-based models are generally strong in

contextual representation, IndoBERT struggled in this domain due to heavy slang usage, creative spellings, and noise

introduced by OCR and ASR transcription. Gambling-specific jargon such as “*scatter*” or “*maxwin*” was often misinterpreted, while metaphorical uses of gambling terms (e.g., “*slot in my heart*”) further reduced classification accuracy. These results indicate that domain-adaptive pretraining or vocabulary augmentation would be necessary for BERT to realize its potential in noisy, multimodal Indonesian datasets.

4.4 Comparative discussion

The comparative evaluation across RNN, CNN, BERT, and AutoML highlights the different strengths and limitations of each approach in detecting online gambling advertisements within Indonesian social media. RNN achieved the highest accuracy (93.07%) and F1-score, reflecting its strong ability to capture sequential dependencies and contextual cues in short, noisy, and code-switched text. This result is consistent with prior studies (e.g., Thapa et al. [16]) that emphasized the effectiveness of recurrent architectures for informal language where subtle cues are spread across sequences.

CNN also performed competitively (92.80%), particularly effective in identifying keyword-rich gambling content. Its success aligns with prior multimodal studies (e.g., Xu et al. [17]) where convolutional filters captured localized textual features. However, CNN’s reliance on local n-grams made it prone to false positives when encountering metaphorical or ambiguous expressions, a limitation similarly reported in hate speech detection tasks.

In contrast, BERT obtained the lowest accuracy (68.12%) despite its theoretical advantage in contextual modeling. This underperformance echoes findings in transformer-based applications such as Adila et al. [19], where heavy noise, slang, and code-switching reduced model effectiveness without domain-specific fine-tuning. Challenges such as creative spellings and OCR/ASR errors in our dataset further weakened BERT’s ability to generalize.

AutoML reached balanced results (90.19%), demonstrating the practicality of automated feature engineering and ensembling for rapid deployment when specialized expertise is limited. Similar to observations by Anitha et al. [25], AutoML showed competitive accuracy but was ultimately constrained by the weaknesses of its strongest base learners, leading to misclassification patterns resembling those of CNN. When compared with prior studies (Table 6), our multimodal-to-text framework shows clear advancements. Unlike Perdana et al. [2], who achieved 96% precision on a Twitter-only dataset using TF-IDF and Random Forest, our approach improves robustness by incorporating multimodal signals (text, image, and audio), enabling generalization across platforms. Similarly, while IndoBERT reported strong results on general NLP tasks [24], our findings highlight the importance of domain adaptation when dealing with illicit, slang-heavy content.

From an application standpoint, three key insights emerge. First, domain-specific data augmentation and noise-robust preprocessing are crucial for handling OCR and ASR imperfections. Second, hybrid architectures that combine RNN/CNN sequential modeling with transformer-based contextual representation hold promise for improved performance. Third, practical deployment, such as real-time moderation of online gambling advertisements, requires balancing accuracy and efficiency—where lightweight models like RNN and CNN remain competitive against heavier

transformer-based solutions.

5. CONCLUSIONS

This study proposed a unified multimodal-to-text framework for detecting online gambling advertisements in Indonesian social media, integrating deep learning architectures (RNN, CNN, and BERT) with an AutoML approach. By incorporating textual content from original posts, OCR-extracted text from images, and speech-to-text transcriptions from videos, the framework effectively addresses the heterogeneous nature of gambling-related promotional content in real-world settings.

Experimental results demonstrate that the RNN model achieved the best overall performance, with an accuracy of 93.07% and strong generalization capability across diverse and noisy inputs. CNN also delivered competitive performance (92.80% accuracy), excelling in detecting keyword-rich promotional language. AutoML achieved 90.19% accuracy, showing the practicality of automated ensembling for rapid deployment in low-resource settings. BERT, while stable in convergence, achieved the lowest accuracy (68.12%), highlighting the challenges of applying transformer-based models without extensive domain-specific adaptation to slang, abbreviations, and noisy text.

Error analysis revealed three primary sources of misclassification: implicit or coded promotional language, noise from OCR and speech-to-text transcription, and contextual ambiguity in metaphorical or non-promotional uses of gambling-related terms. These findings underscore the importance of domain-specific vocabulary augmentation, noise-robust preprocessing, and hybrid modeling approaches for improving detection accuracy in real-world deployments.

Compared to prior studies, our framework not only surpasses traditional machine learning methods but also demonstrates competitive or superior performance relative to other deep learning-based approaches in similar detection domains. The integration of multimodal text extraction with optimized deep learning architectures provides a robust and adaptable solution for detecting illicit promotional content in noisy, mixed-language social media environments.

Beyond research contributions, the framework shows strong potential for real-world application in social media content moderation systems. By standardizing multimodal inputs into a text-based format, it can be embedded into automated pipelines to detect and filter gambling promotions across multiple platforms. Lightweight models such as RNN and CNN support scalable, real-time monitoring with low computational overhead, while hybrid or transformer-augmented designs can be deployed for high-accuracy offline analysis. This flexibility underscores the practical value of the framework for both social media providers and regulators seeking scalable solutions to mitigate the spread of online gambling advertisements.

For future work, this study suggests enhancing transformer-based models through domain-adaptive pretraining to better capture gambling-related slang and creative spelling variations, developing noise-resilient feature learning techniques that address OCR and ASR errors while improving multimodal alignment, and designing hybrid architectures that combine the sequential modeling strengths of RNN or CNN with the contextual representation power of transformers. Addressing these areas can further improve detection accuracy and

robustness, enabling broader applications of the proposed framework in harmful content moderation, digital security, and social media monitoring.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Science, and Technology through the Research and Community Service Program, Fiscal Year 2025 (Grant NO.: 123/C3/DT.05.00/PL/2025 and 145/LL2/DT.05.00/PL/2025). The authors also gratefully acknowledge the support provided by Universitas Bina Darma.

REFERENCES

- [1] Ma'ruf, F., Pattiasina, P.J., Setiawati, R., Camerling, B.C.F., Tuasela, P.E. (2024). The influence of social media usage, internet access, and mobile device penetration on social interaction quality among adolescents in Indonesia. *The Eastasouth Journal of Social Science and Humanities*, 1(3): 106-119. <https://doi.org/10.58812/esssh.v1i03.275>
- [2] Perdana, R.B., Budi, I., Santoso, A.B., Ramadiah, A., Putra, P.K. (2024). Detecting online gambling promotions on Indonesian twitter using text mining algorithm. *International Journal of Advanced Computer Science & Applications*, 15(8): 942-949. <https://doi.org/10.14569/ijacsa.2024.0150893>
- [3] Sangwan, G.D. (2025). Critical study of the financial trends and governance issues in the online gaming and gambling industry. In *Innovative Multidisciplinary Approaches to Global Challenges: Sustainability, Equity, and Ethics in an Interconnected World (IMASEE 2025)*, pp. 224-252. https://doi.org/10.2991/978-2-38476-416-7_11
- [4] Gongane, V.U., Munot, M.V., Anuse, A.D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, 12(1): 129. <https://doi.org/10.1007/s13278-022-00951-3>
- [5] Bassi, D., Fomsgaard, S., Pereira-Fariña, M. (2024). Decoding persuasion: A survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9: 1457433. <https://doi.org/10.3389/fcomm.2024.1457433>
- [6] Suryawati, E., Munandar, D., Riswantini, D., Abka, A.F., Arisal, A. (2018). POS-Tagging for informal language (study in Indonesian tweets). *Journal of Physics: Conference Series*, 971(1): 012055. <https://doi.org/10.1088/1742-6596/971/1/012055>
- [7] Kalaivani, A., Thenmozhi, D. (2025). Composite feature fusion for improved offensive language detection in Tamil social media using MHA-LSTM. *International Journal of Machine Learning and Cybernetics*, 16: 8145-8161. <https://doi.org/10.1007/s13042-025-02715-9>
- [8] Singh, A., Singh, S.K., Chhabra, A., Singh, G., Kumar, S., Arya, V. (2024). Detailed evolution process of CNN-based intrusion detection in the context of network security. In *Digital Forensics and Cyber Crime Investigation*, pp. 70-87.
- [9] Rosid, M.A., Siahaan, D.O., Saikhu, A. (2024). Sarcasm detection in Indonesian-English code-mixed text using multihead attention-based convolutional and bi-directional GRU. *IEEE Access*, 12: 137063-137079. <https://doi.org/10.1109/ACCESS.2024.3436107>
- [10] Wardani, Y.A., Puspitaningtyas, M.O., Ilmi, H.R., Parulian, O.S. (2024). Enhancing text classification performance: A comparative study of RNN and GRU architectures with attention mechanisms. *Journal of Applied Research in Computer Science and Information Systems*, 2(2): 185-190. <https://doi.org/10.61098/jarcis.v2i2.187>
- [11] Yulianti, E., Nissa, N.K. (2024). ABSA of Indonesian customer reviews using IndoBERT: Single-sentence and sentence-pair classification approaches. *Bulletin of Electrical Engineering and Informatics*, 13(5): 3579-3589. <https://doi.org/10.11591/eei.v13i5.8032>
- [12] Baratchi, M., Wang, C., Limmer, S., Van Rijn, J.N., Hoos, H., Bäck, T., Olhofer, M. (2024). Automated machine learning: Past, present and future. *Artificial Intelligence Review*, 57(5): 122. <https://doi.org/10.1007/s10462-024-10726-1>
- [13] Alahmadi, M.D., Alshangiti, M. (2024). Optimizing OCR performance for programming videos: The role of image super-resolution and large language models. *Mathematics*, 12(7): 1036. <https://doi.org/10.3390/math12071036>
- [14] Geneva, D., Shopov, G., Mihov, S. (2019). Building an ASR corpus based on Bulgarian parliament speeches. *Statistical Language and Speech Processing*, 11816: 188-197. https://doi.org/10.1007/978-3-030-31372-2_16
- [15] Bhuva, A.S., Mishra, D. (2025). Gujarati optical character recognition using efficient text feature extraction approaches. *Informatica*, 49(28): 29-58. <https://doi.org/10.31449/inf.v49i28.8341>
- [16] Thapa, S., Adhikari, S., Razzak, I., Lee, R.K.W., Naseem, U. (2025). Hate speech classification in text-embedded images: Integrating ontology, contextual semantics, and vision-language representations. In *Social Networks Analysis and Mining*, pp. 331-342. https://doi.org/10.1007/978-3-031-78538-2_29
- [17] Xu, Q., Du, H., Łukasik, S., Zhu, T., Wang, S., Yu, X. (2025). MDAM³: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, pp. 5285-5296. <https://doi.org/10.1145/3696410.3714498>
- [18] Bell, S., Meglioli, M.C., Richards, M., Sánchez, E., Ropers, C., Wang, S., Williams, A., Sagun, L., Costajussà, M.R. (2025). On the role of speech data in reducing toxicity detection biases. In *Proceedings of the 2025 Conference of the Americas of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1454-1468. <https://doi.org/10.18653/v1/2025.naacl-long.67>
- [19] Adila, A., Lestari, D., Purvarianti, A., Tanaya, D., Azizah, K., Sakti, S. (2024). Enhancing Indonesian automatic speech recognition: Evaluating multilingual models with diverse speech variabilities. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1-6. <https://doi.org/10.1109/O-COCOSDA64382.2024.10800336>
- [20] Mustafa, M.B., Yusoof, M.A., Khalaf, H.K., Rahman Mahmoud Abushariah, A.A., Kiah, M.L.M., Ting, H.N.,

- Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, 12(19): 9541. <https://doi.org/10.3390/app12199541>
- [21] Luckianto, M., Addison, R., Vincent, V., Muliono, Y., Prasetyo, S.Y. (2023). Audio-based Indonesia toxic language classification using bidirectional long short term memory. In 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS), Pangkalpinang, Indonesia, pp. 1-6. <https://doi.org/10.1109/ICORIS60118.2023.10352242>
- [22] Manullang, M.C.T., Rakhman, A.Z., Tantriawan, H., Setiawan, A. (2025). Comparative analysis of CNN, transformers, and traditional ML for classifying online gambling spam comments in Indonesian. *Journal of Applied Informatics and Computing*, 9(3): 592-602. <https://doi.org/10.30871/jaic.v9i3.9468>
- [23] Salehin, I., Islam, M.S., Saha, P., Noman, S.M., Tuni, A., Hasan, M.M., Baten, M.A. (2024). AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, 2(1): 52-81. <https://doi.org/10.1016/j.jiixd.2023.10.002>
- [24] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z.Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*. <https://doi.org/10.48550/arXiv.2009.05387>
- [25] Anitha, M., Dhilipan, J., Kavitha, P.M., Gangadevi, E. (2025). AutoML ecosystem and open-source platforms: Challenges and limitations. In *Automated Machine Learning and Industrial Applications*, pp. 191-205. <https://doi.org/10.1002/9781394272426.ch9>