Deep Learning-Based Detection of Online Gambling Promotion Spam in Indonesian YouTube Comments

Muhammad Zhafran Ammar [1]*, Ricky Eka Putra[2]*, Yuni Yamasari[3]*
* Informatika, Universitas Negeri Surabaya
mzhafranammar@gmail.com [1], rickyeka@unesa.ac.id [2], yuniyamasari@unesa.ac.id [3]

Article Info

Article history:

Received 2025-09-17
Revised 2025-11-10
Accepted 2025-11-15

Keyword:

Online Gambling,
LSTM,
GRU,
Youtube,
Deep Learning.

ABSTRACT

Online gambling promotion has increasingly penetrated social media platforms, with YouTube comments becoming a frequent target for spam-based advertising. Such activities not only violate platform policies but also expose users to harmful content. Addressing this issue requires automated detection systems capable of handling noisy, informal, and highly imbalanced text data. This study investigates the effectiveness of four recurrent neural architectures LSTM, GRU, BiLSTM, and BiGRU for detecting gambling promotion comments in Indonesian YouTube data. To address class imbalance, multiple experimental scenarios were explored, including the original distribution, undersampling, oversampling, and class weighting. Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. The results show that bidirectional models outperformed their unidirectional counterparts, with BiGRU achieving the best overall performance. When combined with class weighting, BiGRU reached 98% accuracy, 0.83 F1-score, and 0.971 ROC-AUC, demonstrating a superior ability to detect minority-class instances. Oversampling improved recall substantially but increased false positives, while undersampling reduced accuracy; class weighting provided the most balanced performance across metrics. These findings confirm that BiGRU with class weighting offers the most practical balance between accuracy, recall, and computational efficiency, making it well-suited for real-time moderation systems. The study provides a strong foundation for future research on transformer-based architectures and cross-platform spam detection in Indonesian social media environments.

I. INTRODUCTION

YouTube, the world's largest video-sharing platform, has transformed the way people consume and engage with online content. As of April 2025, more than 20 million videos are uploaded daily, with the platform already hosting over 20 billion videos in total [1]. With its massive user base, YouTube has evolved into a dynamic ecosystem of content creators and interactive communities, serving not only as a hub for entertainment but also as a medium for information exchange and social interaction. In Indonesia, YouTube is among the most widely accessed platforms, shaping cultural trends and providing a significant channel for communication and marketing. The accessibility of its comment section allows users to express opinions, engage in discussions, and interact directly with creators, making it a vital feature of the platform [2].

The rapid growth of digital technologies has also enabled the expansion of online gambling activities. Online gambling provides individuals with easy access to games and betting platforms, often employing aggressive marketing strategies to attract new users. Psychological factors play a critical role in encouraging participation, as beginners may experience early wins, forming a misperception of high winning probability [3]. Such promotions, including influencer-driven campaigns promising instant profit, reinforce the appeal of online gambling and contribute to its normalization in society [4]. A common tactic employed in online promotion is spamming the comments sections of popular YouTube videos, where promotional messages or links to gambling websites are disseminated with minimal marketing cost [5][6]. Despite YouTube's spam filtering system, these comments often evade detection due to obfuscation tactics such as Unicode characters, Cyrillic alphabets, or special symbols, particularly in comments written in Indonesian [7]. Recent reports further highlight the magnitude of this issue: Google Indonesia disclosed that it blocks around 100,000 online gambling sites every week and removed more than 1.5 million gambling-related ads throughout 2024. Between October 2024 and February 2025 alone, nearly one million pieces of gambling-related content were taken down[8]. Although these numbers do not specifically refer to YouTube comments, they demonstrate the massive scale of online gambling promotion across Google's ecosystem, reinforcing the vulnerability of YouTube's comment section as a target for such illicit activities.

Previous studies have addressed spam detection in YouTube comments using machine learning and Artificial Intelligence (AI). Naive Bayes is popular for its simplicity and effectiveness, achieving accuracies above 80% [7][9]. More advanced methods, such as ensemble techniques, Adaptive Genetic Algorithm (AGA), and XGBoost, improve performance on complex datasets [10][11], while cascaded ensemble models combine multiple classifiers to enhance accuracy [5].

Specifically for online gambling promotions in Indonesian YouTube comments, Samuel and Kristiadi [6] used Natural Language Processing (NLP) and Transformer-based models (IndoBERT) to detect spam, achieving 97% accuracy with balanced precision and recall. Similarly, Riza, et. al. [12] applied Naive Bayes with TF-IDF and preprocessing steps, obtaining 97.1% accuracy. Both studies demonstrate that AI and machine learning can effectively detect and moderate gambling-related spam on YouTube. A recent study by Manullang et. al. [13] compared Support Vector Machine (SVM), Random Forest, Convolutional Neural Network (CNN), IndoBERT, and a lightweight transformer model named Wordformer for Indonesian gambling spam. Wordformer achieved 0.9975 accuracy and macro F1-score, outperforming traditional models while being more computationally efficient than IndoBERT. These studies collectively show that both AI and lightweight deep learning models can effectively detect and moderate gambling-related spam in low-resource language settings such as Indonesian-language.

Despite these advancements, gaps remain in the automatic dDespite these advancements, studies focusing specifically on Indonesian-language gambling spam remain limited compared to English-language datasets, where linguistic resources, labeled corpora, and benchmark datasets are more established. Furthermore, conventional machine learning methods such as Naive Bayes and SVM often struggle to capture semantic context in informal or slang-heavy Indonesian text, and perform poorly under highly imbalanced datasets where non-spam comments dominate. Prior works also tend to focus on experimental accuracy rather than developing scalable or deployable systems capable of handling large volumes of real-world YouTube data. These limitations highlight the need for more robust, context-aware approaches capable of learning complex linguistic patterns in Indonesian-language spam detection.

Recurrent neural networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are especially effective for text classification tasks, as they can capture long-term dependencies in sequences and better represent contextual meaning in noisy, user-generated comments. While LSTM excels at modeling longer dependencies, GRU offers a more computationally efficient alternative with comparable accuracy. In addition, Bidirectional variants of these architectures (BiLSTM and BiGRU) further enhance performance by processing input sequences in both forward and backward directions, thereby capturing past and future context simultaneously. This bidirectional processing is particularly advantageous for detecting gambling-related spam, as promotional cues may appear in different parts of a comment and require full contextual understanding.

To address these research gaps, this study constructs a dataset of Indonesian-language YouTube comments collected via the YouTube API and manually labeled for gambling-related spam detection. It further evaluates and compares the performance of four recurrent neural architectures LSTM, GRU, BiLSTM, and BiGRU to identify the most effective deep learning model for handling informal language and imbalanced spam data. Building on these strengths, this study develops a deep learning–based system for detecting online gambling promotion spam in Indonesian YouTube comments. The findings are expected to contribute not only to academic understanding of spam detection in low-resource languages but also to the development of more effective, automated moderation tools for safer online communities.

II. Method

This chapter describes the methodology applied in the study Deep Learning-Based Detection of Online Gambling Promotion Spam in Indonesian YouTube Comments. The research aims to evaluate the effectiveness of four deep learning models: LSTM, GRU, BiLSTM, and BiGRU for spam classification tasks. To ensure a reliable and reproducible evaluation, a standardized dataset of YouTube comments in Indonesian-language was employed, combined with a consistent preprocessing pipeline and a unified experimental protocol.

The methodology begins with the description of dataset collection and text preprocessing steps, followed by the detailed architecture and configurations of the LSTM, GRU, and their bidirectional variants. Bidirectional models were included to capture contextual information from both past and future tokens, which is crucial for detecting disguised gambling promotions that may appear in different positions within a comment. The chapter concludes with the explanation of training strategies, evaluation metrics, and the computing environment used to validate model performance.

A. Data Collection

This study focuses on user-generated comments from Indonesian YouTube videos as the primary dataset. YouTube was chosen for three main reasons. First, it is one of the most widely accessed platforms in Indonesia, where the comment section serves as a highly interactive space for users to communicate, share opinions, and engage with content creators. Second, the open nature of the comment system makes it vulnerable to misuse by spammers, particularly for promoting online gambling websites at minimal cost. Finally, YouTube comments in Indonesian often include informal language, slang, mixed code-switching, and obfuscation strategies (e.g., Unicode characters, Cyrillic letters, or emojis), making them a challenging yet representative dataset for evaluating spam detection models in low-resource languages.

The dataset was collected using the YouTube Data API v3, capturing both top-level comments and nested replies to preserve conversational context. Each record contained metadata such as comment ID, parent comment ID, author display name, timestamp, and textual content. The initial corpus consisted of 26,208 comments across multiple Indonesian channels, reflecting a natural mixture of entertainment, lifestyle, and popular culture content where gambling spam is commonly disseminated.

A two-stage labeling process was applied to ensure data reliability. First, an automatic pre-labeling step flagged potential gambling spam based on keyword-based heuristics, including brand names (e.g., “pulauwin”, “slot”, “maxwin”), promotional phrases (e.g., “judi online”, “daftar sekarang”, “bonus deposit”), and domain/contact patterns (e.g., URLs, Telegram/WhatsApp handles). Next, manual annotation was performed by two independent annotators who verified each pre-labeled comment and cross-checked ambiguous cases. Disagreements were resolved through discussion to maintain high labeling consistency.

The final dataset was binary-labeled, with 1,400 spam comments (5.34%) and 24,808 non-spam comments (94.66%). This significant imbalance reflects the real-world distribution of spam in social media environments, underscoring the need for robust classification techniques. To address the imbalance during model training, resampling strategies such as SMOTE oversampling and undersampling were applied to the training subset, producing balanced datasets for comparative evaluation.

B. Data Preprocessing

After collection and labeling, all comments underwent a standardized preprocessing pipeline to ensure that the text was clean, consistent, and suitable for input into both deep learning and traditional machine learning models. This step is crucial in social media text analysis because user-generated comments often contain informal language, slang, emojis, URLs, and inconsistent formatting that can hinder model performance.

The preprocessing pipeline included the following steps:

1) *Lowercasing*: All text was converted to lowercase to reduce vocabulary size and normalize casing. This helps the model treat words like “JUDI” and “judi” as the same token.

2) *URL Removal:* URLs and hyperlinks were removed using regular expressions to eliminate non-semantic noise that does not contribute to detecting spam content, for example “daftar di https://slotmaxwin.com sekarang” to “daftar di sekarang”.

3) *Mentions and Hashtags Removal*: Mentions (@username) and hashtags (#promo) were stripped, as they generally do not carry meaningful information for spam detection, as elements like “@admin ayo join #bonusdepo” provide little contextual information and were reduced to “admin ayo join bonusdepo”

4) *Emoji and Symbol Cleaning*: Non-alphanumeric symbols, including emojis and decorative punctuation, were replaced with spaces. This reduces noise while retaining the textual content, so that *“Cuan mantap 🔥 🔥!!!”* became *“cuan mantap”*.

5) *Whitespace Normalization*: Consecutive spaces were collapsed into a single space to maintain consistency for tokenization.

6) *Unicode Normalization*: Text was normalized to NFKD form to standardize accented characters and other Unicode variants, which is important for handling special characters or combined symbols, such as transforming “𝒋𝒖𝒅𝒊 ｏｎｌｉｎｅ” into “judi online”.

7) *Duplicate and Empty Removal*: Duplicate comments and empty entries were removed to prevent bias and redundancy in the dataset.

TABLE I
COMMENT BEFORE AND AFTER PREPROCESSING

Comment	
01:01 Sungguh takjub dengan ❤𝗠𝗢𝗡𝗔 𝟰𝗗🟢! Ceritanya begitu menyentuh, memadukan makna mendalam dengan gameplay yang seru. Visualnya memukau, dan hadiahnya? Benar-benar luar biasa. Wajib coba untuk pengalaman tak terlupakan! 💥	0101 sungguh takjub dengan mona 4d ceritanya begitu menyentuh memadukan makna mendalam dengan gameplay yang seru visualnya memukau dan hadiahnya benarbenar luar biasa wajib coba untuk pengalaman tak terlupakan
Asli deh, ⭐𝗪𝗘𝗧𝗢𝗡𝟴𝟴⭐ sama konten lo sama-sama bikin betah! 🏠	asli deh weton88 sama konten lo samasama bikin betah
06:00 🄿🆄🅻🄰🆄777, menangkan hadiah menarik setiap hari	0600 pulau777 menangkan hadiah menarik setiap hari

Dulu Request ngundang miawaug di kolom komen video bang Radit sama Tara. Akhirnya kejadian 😂	dulu request ngundang miawaug di kolom komen video bang radit sama tara akhirnya kejadian

This sequential preprocessing ensured that the dataset reflected authentic Indonesian social media language while minimizing noise, making it suitable for input into classification models. After these steps, the comments were transformed into numerical representations according to the modeling approach. For the deep learning models (BiLSTM and BiGRU), the text was tokenized using the Keras Tokenizer with a fixed vocabulary size of 10,000, and each sequence was padded to a maximum length of 128 tokens to standardize input dimensions. For traditional machine learning baselines, TF-IDF vectorization was applied to capture both word occurrence and relative importance within the corpus.

C. Model Architectures

In this study, four deep learning architectures were investigated to detect online gambling spam in Indonesian YouTube comments: LSTM, GRU, BiLSTM and BiGRU. These models were chosen because they are widely adopted in text classification tasks and well-suited for handling noisy, sequential data such as YouTube comments. Tokenization, padding, and embedding operations were implemented using the Keras deep learning library.

1) Long Short-Term Memory (LSTM)

LSTM networks are a type of RNN designed to overcome the vanishing gradient problem and capture long-term dependencies in sequential data. LSTM networks maintain an internal cell state that allows them to selectively remember or forget information across long sequences. This makes them highly suitable for processing user-generated comments, where spam-related cues may appear at different positions in a sentence.

LSTM operates through three primary mechanisms: the forget gate, which determines which information from the previous state should be discarded; the input gate, which decides what new information should be stored; and the output gate, which regulates what part of the cell state contributes to the current hidden state. By coordinating these gates, LSTM effectively preserves relevant information while filtering out noise, enabling robust detection of spam phrases that are interspersed with unrelated or informal words in YouTube comments.

2) Gated Recurrent Units (GRU)

Gated Recurrent Units (GRU) are a simplified variant of LSTM that achieve comparable performance with fewer parameters, resulting in faster training. GRUs combine the memory cell and hidden state into a single hidden state, making them more computationally efficient while still capturing long-term dependencies.

GRU relies on two main mechanisms: the update gate, which controls how much of the past information is retained, and the reset gate, which determines how much past information is forgotten when updating the hidden state. This efficient design makes GRU especially suitable for large-scale text classification tasks and effective in handling short, noisy, and informal YouTube comments where spam signals may be sparsely distributed across the text.

3) Bidirectional LSTM (BiLSTM)

The Bidirectional LSTM (BiLSTM) extends the standard LSTM by processing input sequences in both forward and backward directions. This allows each token's representation to be informed by both its preceding and succeeding context.

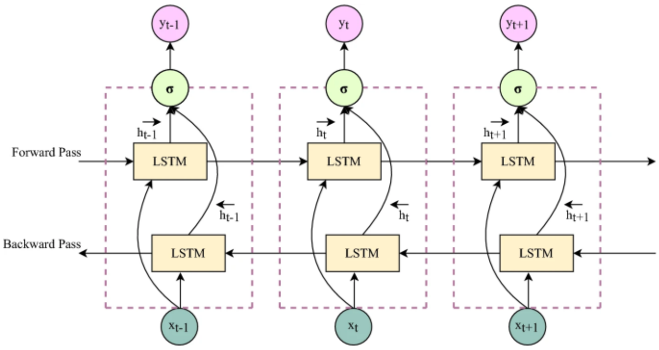

Figure 1. Bi-LSTM layer mechanics[14]

This bidirectional processing is particularly valuable for spam detection in YouTube comments, as promotional cues may appear at the beginning, middle, or end of a sentence, or may even be deliberately obfuscated. By considering both directions, BiLSTM can better identify disguised gambling-related expressions.

4) Bidirectional Gru (BiGRU)

Similar to BiLSTM, the Bidirectional GRU (BiGRU) enhances the standard GRU by processing sequences in both directions. This allows the model to capture contextual cues from both past and future tokens simultaneously, while still maintaining computational efficiency.

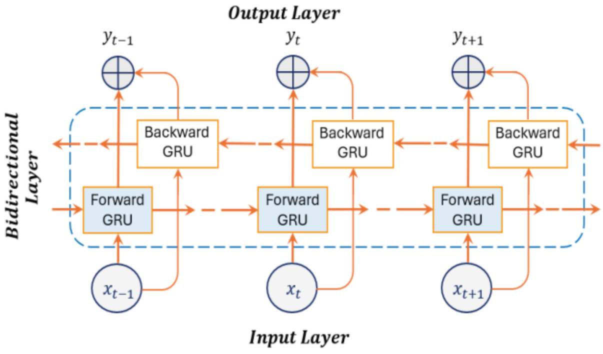

Figure 2. Bi-Gru layer mechanics[15]

BiGRU is especially effective in noisy, user-generated comments, where spam indicators may be scattered across the text. By integrating context from both directions, BiGRU improves the model's ability to detect subtle or shuffled spam patterns in Indonesian YouTube comments.

5) Shared Architecture

All four architectures (LSTM, GRU, BiLSTM, and BiGRU) share a common design framework, ensuring consistent input representation and comparability across models. The first component is the embedding layer, where tokenized words from the preprocessed comments are mapped into dense vector representations of size 100. Word embeddings are widely used in text classification because they enable models to capture semantic and syntactic relationships between words in a continuous space, which is especially useful for handling informal expressions, slang, and obfuscated spellings commonly found in YouTube comments [16][17]. These embeddings are learned during training, allowing the architecture to adapt to the characteristics of Indonesian social media text.

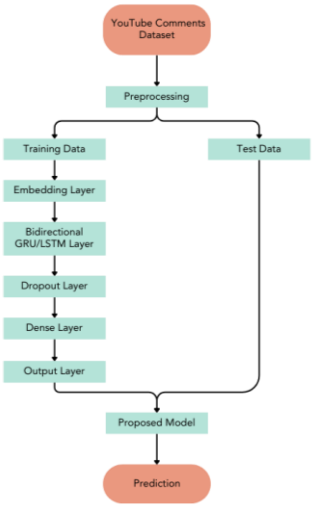

Figure 3. Baseline Model GRU & LSTM

Following the embedding layer, the recurrent layer (either LSTM, GRU, BiLSTM, or BiGRU) processes the sequences to produce context-aware representations for each token. Recurrent architectures are effective in capturing sequential dependencies and contextual cues, which are critical for spam detection tasks where keyword order and context may vary [18][19]. To mitigate overfitting, a dropout layer with a rate of 0.5 is applied after the recurrent layer, as dropout has been proven to be an effective regularization technique in neural networks by preventing co-adaptation of neurons and improving generalization [20].

The output is then passed through a dense hidden layer with 32 units and ReLU activation, enabling non-linear transformations of the learned features. Dense layers allow the model to combine high-level features extracted from sequential processing into a more discriminative representation for classification [21]. Finally, a single output neuron with sigmoid activation produces a probability score for binary classification, distinguishing spam from non-spam comments. The sigmoid activation is commonly used in binary classification problems because it maps the learned features into a probability range between 0 and 1, facilitating threshold-based decision-making. The overall model architecture is illustrated in Figure 3, showing the flow from input preprocessing through embeddings, recurrent processing, dropout, dense transformation, and final classification.

D. Training Data and Experimental Scenarios

The preprocessed dataset was split into training and validation sets using an 80–20 stratified split to maintain the proportion of spam and non-spam comments across both sets. Stratification ensured that both subsets preserved the original distribution of labels, which was highly imbalanced. As shown in Table 1, the training set consisted of 19,846 non-spam comments (94.66%) and 1,120 spam comments (5.34%), while the validation set contained 4,962 non-spam comments (94.66%) and 280 spam comments (5.34%).

TABLE II
LABEL DISTRIBUTION IN TRAINING AND VALIDATION SETS

Dataset	Label	Count	Percentage
Train	0	19,846	94.66%
	1	1,120	5.34%
Test	0	4,962	94.66%
	1	280	5.34%

Because of this imbalance, three experimental scenarios were applied. In the undersampling scenario, the majority class was reduced to match the minority class, producing a perfectly balanced dataset with 1,120 samples each (Table 2). In the oversampling scenario, the minority class was synthetically increased using SMOTE, resulting in 19,846 samples for each class. The original scenario used the natural distribution without modification.

TABLE III
LABEL DISTRIBUTION AFTER BALANCING (TRAINING SET)

Scenario	Label	Count	Percentage
Smote	0	19,846	50%
	1	19,846	50%
Undersample	0	1,120	50%
	1	1,120	50%

To complement these balancing strategies, class weights were also computed on the original dataset and incorporated into the loss function so that misclassification of minority-class samples received a higher penalty. This ensured that the models remained sensitive to spam detection even when trained on imbalanced data.

Figure 4 illustrates the distribution of spam and non-spam comments under the three scenarios: the original data (imbalanced), SMOTE oversampling (balanced with synthetic samples), and undersampling (balanced by reducing majority class). These visualizations highlight the degree of imbalance and the effect of each balancing approach.

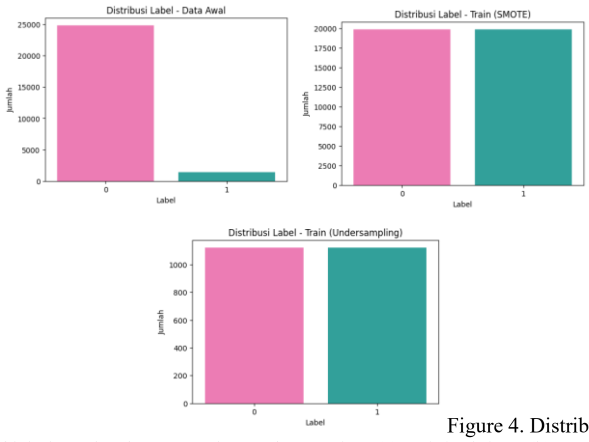

Figure 4. Distribution of labels under three experimental scenarios: (a) Original data, (b) SMOTE oversampling, (c) Undersampling

E. *Model Training*

Model training utilized early stopping by monitoring the validation loss with a patience of 10 epochs, ensuring that training stopped once performance no longer improved. The best model weights were preserved automatically through a model checkpoint. To mitigate overfitting, dropout layers were applied within the architecture, and a class-weighted binary cross-entropy loss was used to address class imbalance, as described earlier.

The experiments were conducted in the Kaggle environment using Python 3.10 and TensorFlow 2.14, supported by an NVIDIA Tesla T4 GPU (16 GB VRAM). All hyperparameters were determined through empirical tuning, where multiple configurations were tested to balance model complexity and training stability. A vocabulary size of 10,000 words was selected to capture diverse Indonesian linguistic variations while preventing excessive sparsity. The maximum sequence length was limited to 128 tokens since most YouTube comments are relatively short, and longer sequences rarely improved performance. The embedding dimension was fixed at 100, which provided sufficient representational capacity without introducing overparameterization.

Each recurrent layer consisted of 64 units, chosen as a trade-off between contextual understanding and computational cost. The dense layer used 32 neurons with ReLU activation to learn higher-level representations before the final sigmoid output, which performed binary classification. Training was carried out with a batch size of 32 for stable gradient updates, and Adam optimizer was used for its adaptive learning capabilities and consistent convergence across all models. The models were trained for a maximum of 50 epochs, with a 20% validation split applied using stratified sampling to preserve class distribution. The ROC–AUC score was used as the primary evaluation metric, as it provides a more reliable assessment of model discrimination ability on imbalanced data compared to accuracy.

F. *Evaluation Model*

Model performance was assessed on the validation set using a combination of quantitative metrics and visualizations. First, predicted labels were generated by applying a threshold of 0.5 to the model's probabilistic outputs. A classification report was then computed to summarize the main metrics for both classes (“Non-Judol” and “Judol”), including precision, recall, and F1-score.

A confusion matrix was generated to provide a detailed breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For better interpretability, the matrix was visualized as a heatmap, showing both raw counts and percentages per class.

To comprehensively evaluate model performance, the following metrics were used on the validation set:

1) *Accuracy:* Proportion of correctly predicted samples

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}$$

2) *Precision*: Ratio of correctly predicted positive observations to all predicted positives

$$Precision=\frac{TP}{TP+FP}$$

3) *Recall (Sensitivity)*: Ratio of correctly predicted positives to all actual positives

$$Recall=\frac{TP}{TP+FN}$$

4) *F1-Score*: Harmonic mean of precision and recall

$$F1-Score=2\times\frac{Precision\times Recall}{Precision+Recall}$$

To evaluate the model's ability to discriminate between spam and non-spam comments, the ROC curve and corresponding AUC score were computed. Probabilistic outputs were used to plot the ROC curve, and cubic spline interpolation was applied to produce a smooth visualization.

Finally, training and validation loss and accuracy curves were plotted across epochs to inspect model convergence and generalization. Cubic spline smoothing was applied to both curves to provide clear visual trends, highlighting whether overfitting or underfitting occurred during training.

III. Results

The evaluation covered four architectures (LSTM, GRU, BiLSTM, BiGRU) under three scenarios (original, undersampling, SMOTE oversampling).

A. *Model Performance Metrics*

Table IV presents a complete comparison across models and settings.

TABLE IV
PERFORMANCE COMPARISON OF MODELS UNDER DIFFERENT DATA BALANCING SCENARIOS

Model	Scenario	Acc	Rec	Pre	F1	ROC
GRU	Original	0.95	0.00	0.00	0.00	0.50
	Undersample	0.05	0.05	1.00	0.10	0.52
	Oversample	0.43	0.08	0.97	0.15	0.92
LSTM	Original	0.95	0.00	0.00	0.00	0.50
	Undersample	0.05	0.05	1.00	0.10	0.40
	Oversample	0.05	0.05	1.00	0.10	0.51
BiGRU	**Original**	**0.98**	**0.82**	**0.85**	**0.83**	**0.97**
	Undersample	0.89	0.32	0.89	0.47	0.96
	Oversample	0.67	0.13	0.92	0.23	0.92
BiLSTM	Original	0.98	0.77	0.89	0.83	0.97
	Undersample	0.85	0.26	0.94	0.40	0.96
	Oversample	0.75	0.17	0.90	0.28	0.92

The results indicate that the bidirectional models (BiLSTM and BiGRU) significantly outperformed their unidirectional counterparts (LSTM and GRU), especially when trained on the original imbalanced dataset. Both BiLSTM and BiGRU achieved accuracy above 98% with high recall and ROC-AUC, demonstrating robustness in handling skewed class distributions without requiring resampling techniques.

Between the two, BiGRU slightly outperformed BiLSTM, particularly in terms of recall (+3%) and ROC-AUC (+0.8%). This shows that BiGRU is more effective at detecting Judol comments, as it captures more minority-class instances while maintaining comparable precision and overall accuracy.

In contrast, the unidirectional LSTM and GRU models performed poorly under the original imbalance, completely failing to detect Judol comments (recall = 0). Although oversampling and undersampling improved recall close to 1.0, this came at the expense of drastically reduced precision and accuracy, making them unsuitable for practical deployment.

Overall, these findings highlight that bidirectional architectures are essential for spam detection tasks on highly imbalanced YouTube comment data, with BiGRU being the most effective and reliable model across different evaluation settings.

B. Confusion Matrix Analysis

The confusion matrices of the BiLSTM and BiGRU models on the original dataset are shown in Figures 5 and 6. Both models achieved high classification accuracy, yet differences can be observed in their handling of the minority (Judol) class. BiGRU yielded higher recall, whereas BiLSTM maintained slightly fewer false positives.

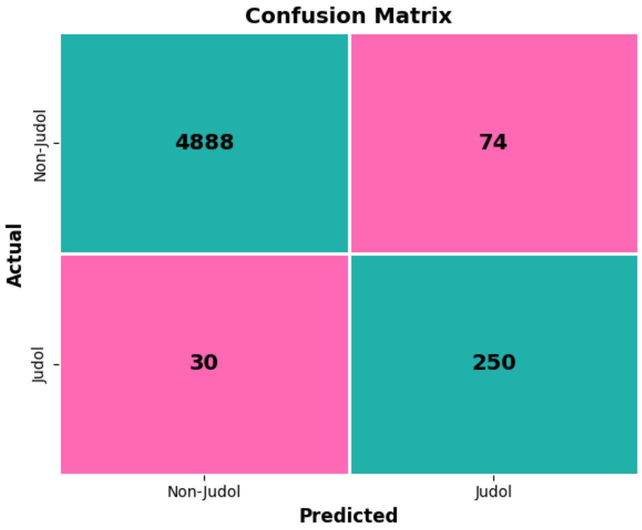

Figure 5. Bi-LSTM Confusion Matrix

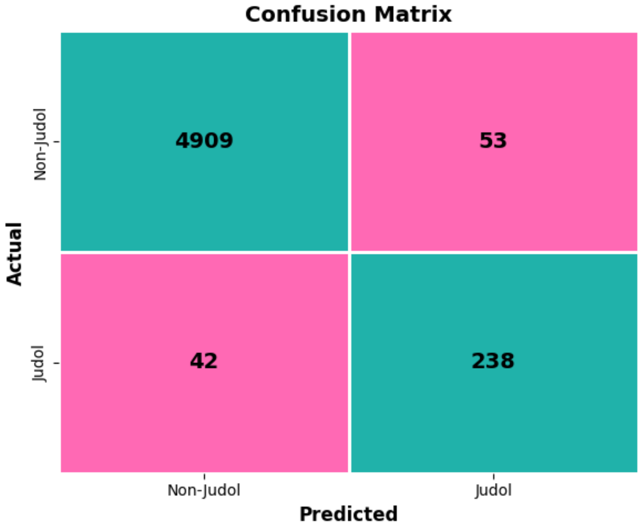

Figure 6. Bi-GRU Confusion Matrix

Bi-LSTM correctly identified 250 Judol comments but misclassified 30 as Non-Judol, while 74 Non-Judol comments were incorrectly predicted as Judol. Bi-GRU performed better, detecting 238 Judol comments with only 42 misclassified as Non-Judol, although it produced 53 false positives. This indicates that BiLSTM is slightly more aggressive in identifying Judol comments, which improves recall but increases false positives compared to BiGru.

For completeness, confusion matrices for all models (LSTM, GRU, BiLSTM, and BiGRU) across the three experimental scenarios (original, oversampling, and undersampling) are provided in Appendix A (Figures A1–A12).

C. Training and Validation Dynamics

Figures 7 and 8 present the training and validation curves of BiLSTM and BiGRU on the original dataset. Both models converged rapidly, surpassing 95% validation accuracy within the first three epochs. While BiLSTM showed minor fluctuations in validation loss after epoch 10, BiGRU maintained more stable validation performance, peaking at 98.5% accuracy.

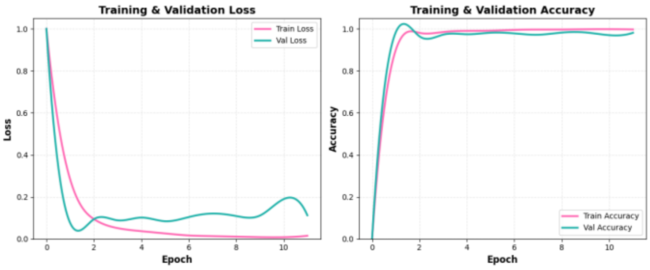

Figure 7. Bi-LSTM Training and Validation

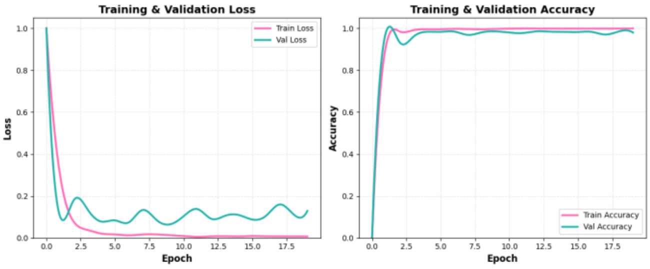

Figure 8. Bi-GRU Training and Validation

The full training and validation curves for all models and experimental scenarios are available in Appendix B (Figures B1–B12).

D. *ROC-AUC Analysis*

The ROC curves of BiLSTM and BiGRU on the original dataset are displayed in Figures 9 and 10.

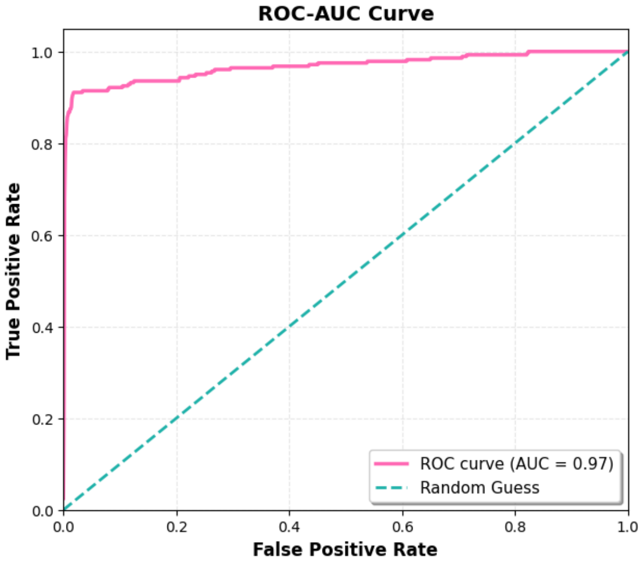

Figure 9. Bi-LSTM ROC-AUC Curve

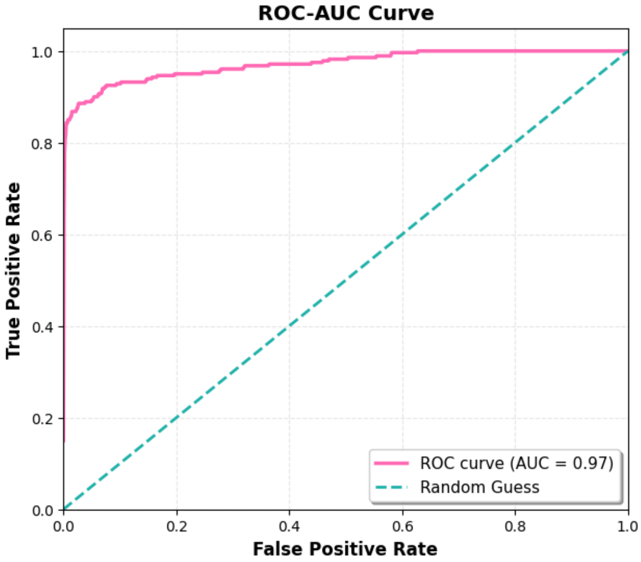

Figure 10. Bi-GRU ROC-AUC Curve

Both achieved ROC-AUC above 0.97, confirming excellent separability between Judol and Non-Judol comments. The BiGRU curve was consistently closer to the top-left corner, reflecting its superior discriminative ability.

IV. DISCUSSIONS

The experimental results demonstrate that both BiLSTM and BiGRU achieved strong performance in detecting online gambling promotion comments in Indonesian YouTube data, with overall accuracy reaching 98 percent. Despite their comparable accuracy, a closer examination of the evaluation metrics shows that BiGRU consistently achieved higher recall and ROC-AUC compared to BiLSTM. Specifically, BiGRU improved recall by 3 percent and ROC-AUC by 0.8 percent, indicating a stronger capability to identify minority-class instances. This is an important finding in the context of spam detection, where the risk of missing harmful promotional content is more critical than incorrectly flagging benign comments.

Another dimension of the analysis involves the different data scenarios applied in this study. Training on the original imbalanced dataset resulted in strong accuracy but relatively weaker recall, as the models were biased toward the majority class. Oversampling with SMOTE improved recall by exposing the models to a larger variety of minority-class samples, though at the cost of increased false positives. Undersampling balanced the classes more directly, producing higher sensitivity but with reduced overall stability due to fewer training samples. The use of class weights provided the most balanced trade-off, ensuring that minority-class misclassifications were penalized during training. These results highlight the importance of handling class imbalance carefully, as it directly impacts a model's ability to detect Judol promotion.

The comparison between unidirectional and bidirectional models further reinforces this conclusion. While both LSTM and GRU delivered competitive results, their bidirectional counterparts consistently achieved stronger performance by leveraging contextual information from both past and future tokens. The improvements were especially notable for BiGRU, which combined the efficiency of GRU with the contextual depth of bidirectionality. In contrast, the unidirectional versions showed limitations in recall, suggesting that access to bidirectional context is crucial when spam-related cues can appear at any position within short, noisy YouTube comments.

This performance advantage can be explained by the complementary strengths of bidirectional processing and the gated recurrent mechanism. Bidirectional learning enables the model to understand the semantic context from both preceding and succeeding words in a comment, which is critical for detecting implicit or fragmented expressions often found in informal YouTube text. Meanwhile, the GRU's gating system effectively regulates information flow without the separate memory cell used in LSTM, making it more efficient in retaining relevant features and discarding noise. The combination of these two mechanisms allows BiGRU to maintain long-term contextual understanding while remaining computationally lightweight, which is particularly advantageous for short, context-limited sequences such as user comments.

Another notable aspect is the computational efficiency of the two models. BiGRU requires fewer parameters and employs a simpler gating mechanism compared to BiLSTM, making it more suitable for real-time applications[22]. In scenarios such as automated comment moderation on social media platforms, computational cost and response time are important considerations. The efficiency advantage of BiGRU therefore provides practical benefits beyond accuracy, especially in large-scale or resource-constrained environments.

The training dynamics of both models also support this conclusion. BiGRU displayed stable validation accuracy and loss throughout training, whereas BiLSTM exhibited slight fluctuations after convergence. This stability suggests that BiGRU is less sensitive to noise in the data and generalizes more effectively. Since YouTube comments are typically short, informal, and highly variable, this robustness is valuable for ensuring reliable performance in practice.

The characteristics of the dataset further explain the observed differences between the two models. Previous studies have shown that GRU-based models are particularly effective for short to medium-length sequences, while LSTM-based models excel in tasks involving long-range dependencies. Given that most of the comments in this dataset are short and context-limited, the structural efficiency of GRU aligns well with the nature of the input, leading to stronger results in minority-class detection.

Despite these encouraging outcomes, several limitations must be acknowledged. The dataset was limited to 19 videos across nine channels, which may introduce topic-specific bias and limit the generalizability of the findings. Furthermore, the study focused exclusively on recurrent neural architectures. Transformer-based models such as IndoBERT or multilingual BERT have been shown to capture complex linguistic features more effectively and may achieve even stronger results in similar tasks. Future research could expand the dataset across multiple platforms, incorporate additional deep learning architectures such as attention-augmented GRU or hybrid ensembles, and evaluate model performance in cross-domain scenarios.

Overall, the findings confirm that BiGRU provides a practical balance between accuracy, recall, and computational efficiency. These characteristics make it a strong candidate for real-world deployment in automated moderation systems aimed at detecting online gambling promotion spam in Indonesian social media environments.

V. Conclusion

This study investigated the detection of online gambling promotion (Judol) comments in Indonesian YouTube data using four recurrent neural architectures: LSTM, GRU, BiLSTM, and BiGRU. Multiple data-balancing strategies were also explored, including the original imbalanced dataset, undersampling, oversampling, and class weighting, to evaluate their effect on minority-class detection.

The results show that the bidirectional variants (BiLSTM and BiGRU) consistently outperformed their unidirectional counterparts, with BiGRU achieving the strongest overall performance. Specifically, BiGRU trained on the original dataset with class weighting achieved the highest balance across metrics, reaching 98.19% accuracy, 0.83 F1-score, and 0.971 ROC-AUC. Compared to BiLSTM, BiGRU improved recall by approximately 3% and ROC-AUC by nearly 1%, demonstrating a stronger ability to capture minority-class (Judol) instances.

The experiments with different data scenarios also provide key insights. Training on the original imbalanced dataset yielded high accuracy but low recall, as models favored the majority class. Oversampling improved recall substantially (up to 0.92 for Judol) but reduced overall accuracy due to increased false positives. Undersampling allowed models to detect Judol more aggressively but significantly harmed overall accuracy. Class weighting emerged as the most balanced strategy, allowing both BiLSTM and BiGRU to generalize effectively without sacrificing accuracy.

These findings confirm that BiGRU, when combined with class weighting, represents the most practical approach for detecting Judol promotion in YouTube comments. Its superior recall, strong discriminative ability, and computational efficiency make it suitable for real-time moderation systems where failing to detect harmful content carries greater risk than false positives.

Although the dataset size and scope were limited, the findings provide a strong foundation for future work. Expanding the dataset, incorporating transformer-based models, and exploring hybrid architectures could further enhance detection performance.

In conclusion, this study demonstrates that BiGRU with class weighting provides the optimal balance of accuracy (98%), recall (0.85–0.89), and ROC-AUC (>0.97), establishing it as a promising model for real-world deployment in automated moderation systems aimed at combating online gambling promotion spam in Indonesian social media environments.

DAFTAR PUSTAKA

[1] YouTube, "YouTube Press Statistics." [Online]. Available: https://www.youtube.com/yt/about/press/

[2] A. S. Xiao and Q. Liang, "Spam detection for Youtube video comments using machine learning approaches," *Mach. Learn. with Appl.*, vol. 16, no. December 2023, p. 100550, 2024, doi: 10.1016/j.mlwa.2024.100550.

[3] A. A. Makarin and L. Astuti, "Faktor yang Mempengaruhi Mahasiswa Melakukan Perjudian Online," *Indones. J. Crim. Law Criminol.*, vol. 3, no. 3, pp. 180–189, 2023, doi: 10.18196/ijclc.v3i3.17674.

[4] Pande Putu Rastika Paramartha, Anak Agung Sagung Laksmi Dewi, and I Putu Gede Seputra, "Sanksi Pidana terhadap Para Pemasang dan Promosi Iklan Bermuatan Konten Judi Online," *J. Prefer. Huk.*, vol. 2, no. 1, pp. 156–160, 2021, doi: 10.22225/jph.2.1.3062.156-160.

[5] H. Oh, "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model," *IEEE Access*, vol. 9, pp. 144121–144128, 2021, doi: 10.1109/ACCESS.2021.3121508.

[6] Samuel and D. Prasetya Kristiadi, "Deteksi Teks Promosi Judi Online Menggunakan Ai Dengan Kombinasi NLP Dan Deep Learning," *J. Sist. Inf. dan Teknol.* , vol. 5, no. 2 SE-Artikel, pp. 179–185, Jul. 2025, doi: 10.56995/sintek.v5i2.179.

[7] J. R. Fernando, R. Budiraharjo, and E. Haganusa, "Spam Classification on 2019 Indonesian President Election Youtube Comments Using Multinomial Naïve-Bayes," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 1, pp. 37–44, 2019, doi: 10.24014/ijaidm.v2i1.6445.

[8] Bloomberg Technoz, "Konten Judi Online Menjamur di Komen YouTube, Google Menjawab," Feb. 2025. [Online]. Available: https://www.bloombergtechnoz.com/detail-news/63330/konten-judi-online-menjamur-di-komen-youtube-google-menjawab/2

[9] N. M. Samsudin, C. F. B. Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, "Youtube spam detection framework using naïve bayes and logistic regression," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1508–1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[10] A. O. Abdullah, M. A. Ali, M. Karabatak, and A. Sengur, "A comparative analysis of common YouTube comment spam filtering techniques," *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ISDFS.2018.8355315.

[11] N. Ghatasheh, I. Altaharwa, and K. Aldebei, "Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction," *IEEE Access*, vol. 10, no. July, pp. 84365–84383, 2022, doi: 10.1109/ACCESS.2022.3196905.

[12] F. Jauhari, M. Riza, R. G. Guntara, and M. R. Nugraha, "Indonesian Journal of Digital Business Implementasi Algoritma Naive Bayes untuk Filtrasi Spam Komentar Judi Online pada YouTube," vol. 5, no. 2, pp. 411–423, 2025.

[13] M. C. T. Manullang, A. Z. Rakhman, H. Tantriawan, and A. Setiawan, "Comparative Analysis of CNN, Transformers, and Traditional ML for Classifying Online Gambling Spam Comments in Indonesian," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 592–602, 2025, doi: 10.30871/jaic.v9i3.9468.

[14] D. Naik and C. D. Jaidhar, "A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM," *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00664-6.

[15] E. Mahdi, C. Martin-Barreiro, and X. Cabezas, "A Novel Hybrid Approach Using an Attention-Based Transformer + GRU Model for Predicting Cryptocurrency Prices," *Mathematics*, vol. 13, no. 9, pp. 1–19, 2025, doi: 10.3390/math13091484.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.3781

[17] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746–1751, 2014, doi: 10.3115/v1/d14-1181.

[18] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.

[19] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[21] I. Goodfellow, "Front Matter," *Linear Algebr.*, pp. i–ii, 2014, doi: 10.1016/b978-0-12-391420-0.09987-x.

[22] M. J. Hamayel and A. Y. Owda, "A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms," *AI*, vol. 2, no. 4, pp. 477–496, 2021, doi: 10.3390/ai2040030.