Jurnal Statistika Dan Komputasi (STATKOM), Vol. 5 No. 1 (June 2026)
ISSN : 2963-038X (print), ISSN : 2963-0398
DOI : https://doi.org/10.32665/statkom.v5i1.6412

Multimodal Deep Learning for Online Gambling Promotion Detection on Indonesian Social Media

Danish Rafie Ekaputra[1], Muhamad Syukron[2]

[1, 2] Statistics Department, Institut Teknologi Sepuluh Nopember
E-mail: muhamad.syukron@its.ac.id

Submitted: March 31, 2026 ***Revised:*** June 9, 2026 ***Accepted:*** June 15, 2026

Abstract

Background: *Online gambling promotion has become a major social problem in Indonesia, generating billions of rupiah in annual transactions despite strict legal prohibitions. Existing detection methods primarily focus on textual content while overlooking the visual information commonly used in social media promotions.*
Objective: *This study proposes a multimodal deep learning approach that combines image and text representations for detecting online gambling promotions on Indonesian social media and introduces the first publicly available multimodal dataset for this task.*
Methods: *A dataset of 5,028 labeled image-caption pairs was collected from Facebook, TikTok, and X, comprising 2,195 gambling promotion and 2,833 non-promotion samples. Three image embedding models, three text embedding models, and six multimodal fusion strategies were evaluated using a lightweight Multilayer Perceptron classifier.*
Results: *SigLIP 2 achieved the best image-only performance, while Multilingual E5-large achieved the best text-only performance. The contrastive similarity fusion strategy achieved the highest overall performance, with 97.51% accuracy and a 97.18% F1-score.*
Conclusion: *The proposed multimodal approach effectively detects online gambling promotions by leveraging complementary visual and textual information. The publicly available dataset also provides a benchmark to support future research.*

Keywords : *Multimodal Classification, Online Gambling Detection, Text Embedding, Image Embedding, Indonesian Social Media.*

Abstrak

Latar Belakang: Promosi judi daring telah menjadi permasalahan sosial yang serius di Indonesia dengan nilai transaksi mencapai miliaran rupiah setiap tahun meskipun telah dilarang secara hukum. Sebagian besar metode deteksi masih berfokus pada analisis teks dan mengabaikan informasi visual pada media sosial.
Tujuan: Penelitian ini mengusulkan pendekatan pembelajaran mendalam multimodal yang menggabungkan representasi citra dan teks untuk mendeteksi promosi judi daring serta memperkenalkan dataset multimodal pertama yang tersedia secara publik untuk tugas ini.
Metode: Dataset terdiri atas 5.028 pasangan citra dan caption dari Facebook, TikTok, dan X, mencakup 2.195 sampel promosi judi daring dan 2.833 sampel nonpromosi. Tiga model embedding citra, tiga model embedding teks, dan enam strategi fusi multimodal dievaluasi menggunakan pengklasifikasi Multilayer Perceptron.
Hasil: SigLIP 2 memberikan kinerja terbaik pada klasifikasi citra, sedangkan Multilingual E5-large unggul pada klasifikasi teks. Strategi fusi contrastive similarity mencapai akurasi 97,51% dan F1-score 97,18%.
Kesimpulan: Pendekatan multimodal yang diusulkan efektif memanfaatkan informasi visual dan tekstual untuk mendeteksi promosi judi daring. Dataset yang dipublikasikan juga menyediakan tolok ukur bagi penelitian selanjutnya.

Kata kunci: Klasifikasi Multimodal, Deteksi Judi Online, Embedding Teks, Embedding Gambar, Media Sosial Indonesia.

INTRODUCTION

More than 80% of countries worldwide permit some form of gambling, with higher prevalence in Europe and the Americas (Ukhova et al., 2024). In Asia, the proportion is lower, at approximately 61%. Despite its legal status in many regions, recent studies indicate a shift toward stricter regulation. Ukhova et al. (2024) report that 67 European countries, representing about 35% of the region, have implemented tighter gambling regulations in response to concerns about negative impacts. Fisher et al. (2025) show that gambling causes harms affecting both individuals and society, including financial difficulties, relationship problems, and adverse mental and physical health outcomes. Muggleton et al. (2021) also report that gambling is associated with higher risks of future unemployment and physical disability and, at severe levels, increased mortality.

In Indonesia, gambling is illegal under Article 303 of the Criminal Code (KUHP) (Junaedi, 2024). This law prohibits offering gambling opportunities, organizing gambling activities for profit, and providing public access to such activities. Violations are considered criminal offenses and may result in imprisonment or fines. Even so, gambling continues to occur clandestinely, ranging from traditional activities such as cockfighting and sports betting to the increasingly prevalent online gambling.

Currently, online gambling constitutes one of the most significant economic challenges in Indonesia. Data from the Financial Transaction Reports and Analysis Center (PPATK) indicate that approximately 76 % of online gamblers belong to low-income households (Kongah, 2025). Many of these individuals rely on government social assistance intended for essential needs, yet a substantial portion of these funds is diverted to online gambling. In 2024 alone, the total value of online gambling transactions among this group reached IDR 957 billion (approximately USD 58.9 million) (Ilham, 2025). Alarmingly, in the first quarter of 2025, children aged 10–16 were reported to be addicted to online gambling, with total transactions amounting to IDR 2.2 billion (USD 140,000) (Kongah, 2025). These activities not only result in considerable financial losses but are also associated with mental health issues, increased criminal activity, and a rise in divorce rates (Santika, 2025).

Given the numerous drawbacks associated with online gambling, the Indonesian government has attempted to block online gambling websites and accounts that promote them (Rahman, 2025). However, such promotions are often difficult to detect because the messages are typically implicit, embedded in social media captions. In addition, online gambling logos are frequently disguised, and their names do not explicitly reference gambling. Visual elements may also be small or subtly placed within posts. Consequently, effective detection requires analyzing both the textual captions and the images within social media content.

Indonesia has a large and highly active social media ecosystem across major platforms such as TikTok, Facebook, and X. This broad platform coverage makes public social media a relevant environment for observing and monitoring online gambling promotion. Because promotional content can spread across multiple platforms and appears at the level of individual posts, manual monitoring is highly time-consuming. Therefore, automated detection systems are needed, and artificial intelligence-based models offer an effective approach to identify and reduce such content.

Machine learning methods have been widely applied for classification tasks in Indonesian applied-statistics research, including Naive Bayes-based classification (Arifat et al., 2023) and sentiment analysis of user reviews (Ginasputri et al., 2025). Several

studies have examined online gambling in Indonesia. Some focused on YouTube comments (Pratama et al., 2025; Kamdan et al., 2025; Widiyanto et al., 2025; Manullang et al., 2025), while others analyzed tweets on X (Perdana et al., 2024). However, these studies addressed only a single modality, namely text, even though online gambling promotions often appear in images within social media posts. Other studies, Maldini et al. (2025), Alfiansyah & Muzaki (2025), Muzakir et al. (2025), incorporated multiple media types such as images, text, and audio, but they relied on extracting text from images and audio using Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR). Consequently, these approaches perform text classification across multiple sources rather than true multimodal classification. Recent studies have further explored Indonesian gambling content detection through BERT-based systems (Situmorang et al., 2025), convolutional neural networks (Cahyo et al., 2025), and comparisons of deep learning classifiers (Liem et al., 2025), yet all remain limited to text-only analysis.

In related content moderation domains, multimodal fusion of text and image features has demonstrated substantial improvements for hate speech detection (Prabhu & Seethalakshmi, 2025), fake news detection (Lu & Yao, 2025; Lin et al., 2024), and hybrid text-image classification (Gupta & Kishan, 2025), while Zhang et al. (2025) applied a fusion-assisted multimodal large language model specifically for gambling website identification. Studies on text embedding evaluation for Indonesian natural language processing (NLP) (Sutriawan et al., 2025) and social media toxicity classification (Fan et al., 2021) further inform the selection of effective text representations for low-resource language settings.

From the reviewed literature, it is evident that, both in Indonesia and globally, no study has yet applied multimodal fusion for detecting online gambling promotions. Furthermore, there is a lack of publicly available multimodal datasets. To address this gap, this study also creates a multimodal dataset, which can be used by other researchers for future investigations and development in this area.

METHOD

Research Design

This study adopts a three-stage pipeline for multimodal gambling promotion detection, as illustrated in Figure 1. In the first two stages, image and text embeddings are extracted independently using frozen pretrained encoders, allowing each modality to be evaluated in isolation. In the third stage, the best-performing image encoder (SigLIP 2 (Tschannen et al., 2025; Chen et al., 2023)) and text encoder (Multilingual E5-large (Wang et al., 2024; Artetxe & Schwenk, 2019)) are combined through six fusion strategies spanning three paradigms: early fusion (concatenation), intermediate fusion (projection, contrastive similarity), and late fusion (average, weighted, stacking). A lightweight MLP classifier is trained on the resulting representations for binary classification.

This modular design serves three purposes. First, evaluating encoders independently before fusion enables fair comparison of their representational quality. Second, the decoupled architecture allows each component to be analyzed and replaced without retraining the entire system. Third, the dataset size of 5,028 samples is insufficient for fine-tuning encoders with hundreds of millions of parameters; using frozen pretrained encoders paired with a lightweight MLP classifier (approximately 17K trainable parameters) mitigates overfitting while leveraging the rich representations learned during large-scale pretraining. The training follows a two-phase protocol. In Phase 1, the classifier is trained

on the training set (70%) with early stopping based on validation loss (10%), enabling hyperparameter selection without test set exposure. In Phase 2, the classifier is retrained from scratch on the combined training and validation sets (80%) and evaluated on the held-out test set (20%). This protocol maximizes training data utilization while preserving rigorous evaluation on unseen samples. Unlike prior studies on Indonesian gambling detection (Maldini et al., 2025; Alfiansyah & Muzakir, 2025; Muzakir et al., 2025), which incorporated multiple media types but relied on OCR and ASR to extract text from images and audio, effectively performing text-only classification across multiple sources, this work employs separate vision and language encoders that preserve modality-specific semantics and fuse them through explicit multimodal strategies. To the best of our knowledge, this is the first study to systematically compare multiple fusion paradigms for online gambling promotion detection and to provide a publicly available multimodal dataset for this task.

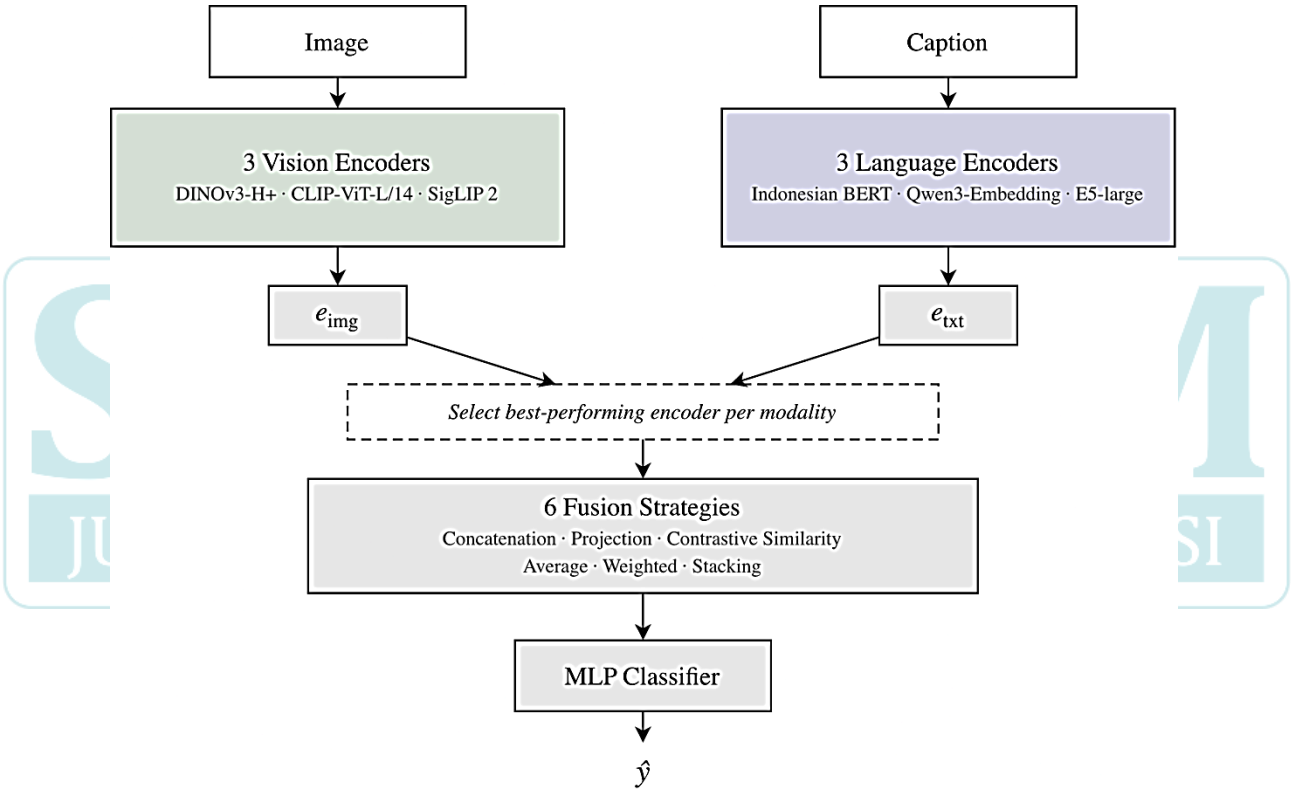


Figure 1. Experimental pipeline for multimodal gambling promotion detection using three vision encoders, three language encoders, and six fusion strategies.

Population and Sample

The population of this study comprises all public social media posts related to online gambling on Facebook, TikTok, and X in Indonesia. The sample consists of 5,028 posts collected through keyword-based scraping from November to December 2025, representing three major social media platforms with high user activity in the country.

Sampling Techniques

This study employs purposive sampling through keyword-based web scraping. Posts were retrieved from Facebook, TikTok, and X using the keywords “Judi Online,” “Judol,” and “Slot Gacor,” yielding between 1,419 and 2,167 samples per platform after deduplication and filtering to remove duplicate posts and irrelevant content.

Research Subjects

Both the images and text captions from each post were collected and manually labeled at the post level. A label of 1 indicates

an online gambling promotion, while a label of 0 indicates a non-promotion. The labeling protocol followed predefined operational criteria: posts were labeled as gambling promotion when they contained promotional cues such as gambling platform names, betting invitations, bonus offers, or gambling-related hashtags, whereas posts that mentioned gambling without promotional intent were labeled as non-promotion. Ambiguous samples were reviewed against these operational criteria before finalizing the labels. Because labeling was conducted by a single annotator, inter-annotator agreement was not computed; therefore, labeling subjectivity is acknowledged as a limitation of the dataset construction process.

To address ethical considerations in social media data collection, this study used only publicly accessible posts retrieved through keyword-based scraping, and the analysis was conducted at the post level rather than for user identification or profiling. Personal information such as names, usernames, account identifiers, phone numbers, personal locations, and other sensitive metadata was not collected, analyzed, or reported. The final dataset contains 5,028 samples, with 2,195 gambling promotion posts (43.7%) and 2,833 non-gambling posts (56.3%). Table 1 presents the distribution of samples across platforms and classes to make potential platform and class imbalance transparent; this transparency supports bias awareness, although keyword-based sampling and single-annotator labeling may still introduce selection and labeling bias. Table 2 shows representative examples from each class, including a gambling promotion post with explicit bonus offers and a non-promotion post that mentions "judol" in the context of reporting its negative consequences. The dataset is publicly available at https://huggingface.co/datasets/0xRafie/indonesian-gambling-multimodal-dataset.

Table 1. Dataset Distribution across Platforms and Classes.

Platform	Gambling	Non-Gambling	Total
Facebook	927	492	1,419
TikTok	199	1,243	1,442
X	1,069	1,098	2,167
Total	2,195	2,833	5,028

Table 2. Examples of Gambling Promotion and Non-Promotion Posts.

ID	Image	Caption	Label
1		Hai Bos Ku, 77lucky bagi-bagi hadiah ni. ... 100 juta rupiah. Buruan!! #77lucky #judionlin	1
2		Pria terjun bebas gara gara judol!!	0

Data Analysis Techniques

This study employs three vision models for image embedding extraction: DINOv3 (Siméoni et al. 2025; Uelwer et al. 2025), CLIP-ViT-L/14 (Radford et al., 2021), and SigLIP 2 (Tschannen et al. 2025; Chen et al. 2023). These models represent distinct paradigms in visual representation learning: self-supervised learning (DINOv3) and vision-language contrastive pretraining (CLIP and SigLIP 2), enabling a comparison of their suitability for downstream classification tasks.

DINOv3, developed by Meta AI (Siméoni et al., 2025), is included as a self-supervised vision foundation model to evaluate whether visual representations learned without language supervision are effective for online gambling promotion detection. Unlike vision-language models such as CLIP and SigLIP 2, DINOv3 learns visual representations from image data through self-supervised training,

making it useful as a comparison point for assessing the contribution of language-aligned visual pretraining. For this study, the ViT-H/16+ variant is used as a frozen image encoder and extracts 1280-dimensional image embeddings from each image. The model is relevant to this task because gambling promotional images may contain visual cues such as platform logos, betting interfaces, promotional banners, and bonus advertisements. These features can appear even when the accompanying text caption is ambiguous, making image-based representations important for comparison with text and multimodal models. For the detail architecture, see the Figure 2.

Contrastive Language-Image Pretraining (CLIP), developed by OpenAI (Radford et al., 2021), is used as a vision-language contrastive model for extracting image embeddings. CLIP learns visual representations by aligning images with natural language descriptions during pretraining, which makes it suitable for tasks where visual content has semantic relationships with textual concepts. In this study, CLIP serves as a contrastive vision-language baseline against the self-supervised DINOv3 model.

For the experiment, we use the ViT-L/14 variant at 336×336 input resolution and extract 768-dimensional projected image embeddings. The model is used only as a frozen feature extractor, and the extracted embeddings are passed to the downstream MLP classifier. This setup allows the study to evaluate whether language-aligned visual features improve the detection of gambling promotional content compared with purely visual self-supervised representations. Figure 3 summarizes the contrastive pretraining concept.

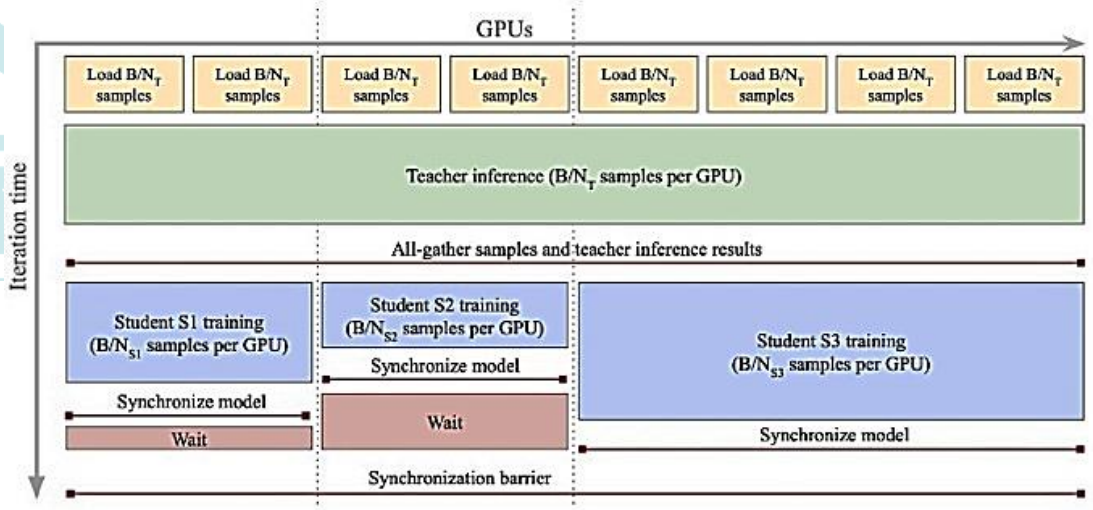


Figure 2. Multi-student distillation procedure in DINOv3 (Siméoni et al., 2025) for self-supervised visual representation learning. This study uses the ViT-H/16+ variant.

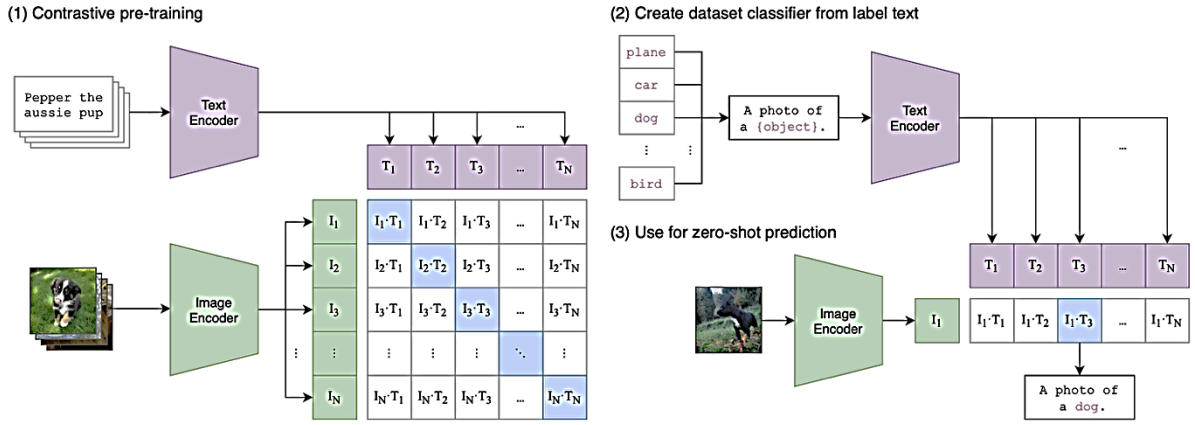


Figure 3. CLIP contrastive pre-training approach for jointly learning image and text representations (Radford et al., 2021). This study uses the ViT-L/14 variant.

A multilingual vision-language encoder developed by Google (Tschannen et al., 2025). This model extends contrastive image-text representation learning to a broader multilingual setting. Compared with CLIP, SigLIP 2 is designed to improve semantic understanding and visual localization through its updated training recipe, making it relevant for detecting promotional content that combines visual symbols, text-like graphics, and culturally specific advertisement patterns. For this study, the So400m variant at 384×384 input resolution is used and extracts 1152-dimensional embeddings through its Multihead Attention Pooling (MAP) head. The model is used as a frozen image encoder to ensure that performance differences reflect embedding quality rather than encoder fine-tuning. SigLIP 2 is particularly relevant because Indonesian gambling promotion posts often contain visual elements whose meaning is connected to promotional language, such as bonuses, platform branding, and betting-related visual layouts. Figure 4 summarizes the SigLIP 2 training recipe.

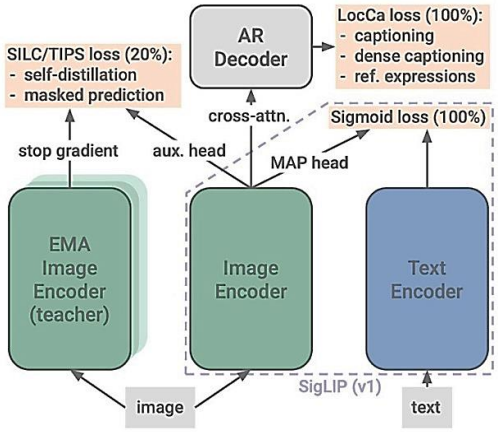


Figure 4. SigLIP 2 training recipe using sigmoid-based contrastive loss (Tschannen et al., 2025). This study uses the So400m variant at 384×384 resolution.

This study employs three text embedding models for extracting semantic representations from Indonesian social media captions: Indonesian BERT (Cahya) (Wirawan, 2022; Imaduddin et al., 2023), Qwen3-Embedding-0.6B (Zhang et al., 2025; Stankevičius & Lukoševičius, 2024), and Multilingual E5-large (Wang et al., 2024; Artetxe & Schwenk, 2019). These models represent different paradigms: monolingual Indonesian pre-training (Cahya BERT), decoder-based LLM embedding (Qwen3), and multilingual contrastive training (E5), allowing us to evaluate the trade-offs between language-specific depth and multilingual breadth.

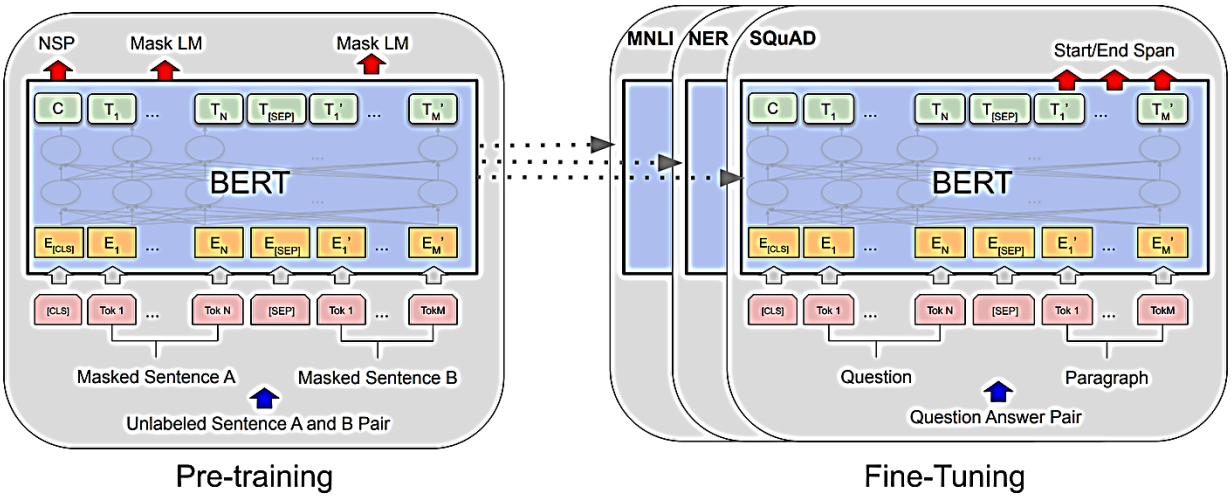


Figure 5. BERT pre-training and fine-tuning procedures (Devlin et al., 2019). Indonesian BERT (Cahya) follows this procedure.

To evaluate the effectiveness of a monolingual Indonesian language model, we include BERT-base-Indonesian-1.5G (Wirawan, 2022). This community-developed model was pre-trained exclusively on Indonesian text. The model follows the standard BERT-base architecture with 12 transformer layers, 12 attention heads, and a hidden dimension of 768 (See Fig. 5). The model was pre-trained on approximately 1.5 GB of Indonesian text, comprising Indonesian Wikipedia (522 MB) and Indonesian newspaper articles from 2018 (1 GB), using the Masked Language Modeling (MLM) objective. The vocabulary consists of 32,000 WordPiece tokens optimized for Indonesian text. Unlike the multilingual models E5-large and Qwen3- Embedding that support 100 or more languages, this model is trained solely on Indonesian data. Including this monolingual baseline allows users to assess whether language-specific pre-training on a smaller corpus provides advantages over multilingual models with broader but potentially less focused Indonesian coverage. Extract 768-dimensional embeddings using mean pooling over the last hidden states.

Qwen3-Embedding is a text embedding model developed by Alibaba's Qwen team (Zhang et al., 2025). Unlike BERT and E5, which are encoder-based models, Qwen3-Embedding is built from a large language model backbone and is designed to produce general-purpose text embeddings across multiple languages and domains. For this study, the 0.6B parameter variant will be used to balance embedding quality and computational efficiency. Indonesian captions are processed using a consistent extraction setting without task-specific prompts, so that the comparison with other text models remains fair. Embeddings are extracted from the last token position and then L2-normalized, producing 1024-dimensional vectors. This model is included to evaluate whether decoder-based multilingual embeddings provide advantages for informal Indonesian social media captions containing slang, abbreviations, and gambling-related promotional terms. Figure 6 illustrates the last-token pooling approach used by Qwen3-Embedding.

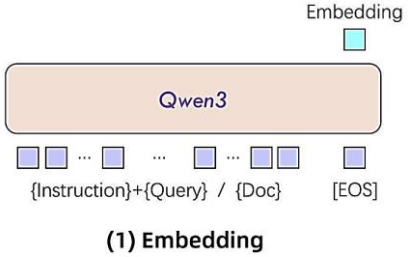


Figure 6. Qwen3-Embedding architecture with last-token pooling (Y. Zhang et al., 2025). This study uses the 0.6B variant.

Multilingual E5-large is a text embedding model developed by Microsoft Research (Wang et al., 2024). The model follows a two-stage training procedure: contrastive pretraining on 1 billion multilingual text pairs, followed by finetuning on a combination of labeled datasets. The "E5" name derives from "EmbEddings from bidirEctional Encoder rEpresentations," reflecting its foundation on transformer encoder architectures.

The model is built upon the XLM-RoBERTa architecture (Conneau et al., 2020) with 24 transformer layers and produces embeddings with 1024 dimensions. It supports 100 languages, including Indonesian, making it particularly suitable for processing Indonesian social media content. The training data encompasses diverse text pairs from multiple languages, enabling robust cross-lingual transfer capabilities.

For embedding extraction, the model applies mean pooling over the last hidden states, followed by L2 normalization. Following the E5 protocol, query texts are prefixed with "query: " to distinguish them from document texts during retrieval tasks. The model's strong performance on the MTEB multilingual benchmark demonstrates its effectiveness for cross-lingual text understanding.

For classification, this study employs a lightweight Multilayer Perceptron (MLP) as the classifier head. The MLP architecture is kept simple and consistent across all embedding models to ensure fair comparison of the embedding representations themselves, rather than classifier complexity.

Table 3 summarizes the embedding configuration for each model. The embedding dimension, pooling strategy, and input size vary across models, as each model was designed with a different architecture and pretraining objective. For image models, DINOv3 uses CLS token pooling at 224×224 resolution, CLIP uses a linear projection head at 336×336, and SigLIP 2 uses Multihead Attention Pooling (MAP) at 384×384. For text models, Indonesian BERT and E5-large use mean pooling over hidden states, while Qwen3-Embedding uses last-token pooling as it is a decoder-based model. E5-large and Qwen3-Embedding additionally apply L2 normalization, and E5-large prepends a "query: " prefix to input texts following its training protocol.

Table 3. Embedding Configuration for Each Model.

Model	Dim	Pooling	Input Size
Image Models			
DINOv3-H+	1280	CLS token	224x224
CLIP-ViT-L/14	768	Projection	336x336
SigLIP 2	1152	MAP head	384x384
Text Models			
Indonesian BERT	768	Mean	512 tokens
Qwen3-Embedding	1024	Last token + L2	512 tokens
E5-large	1024	Mean + L2	512 tokens

The classifier consists of two fully connected layers. The first layer projects the input embedding to a hidden dimension of 128 units, followed by a ReLU activation function and a dropout layer with probability 0.3 for regularization. The second layer maps the hidden representation to a single output unit for binary classification. The final prediction is obtained by applying a sigmoid function to the output logit.

The model is trained using Binary Cross-Entropy loss with the Adam optimizer at a learning rate of $1 \times 10{-}3$. Training is conducted for a maximum of 50 epochs with early stopping based on validation loss, using a patience of 5 epochs. A learning rate scheduler with the ReduceLROnPlateau strategy reduces the learning rate by a factor of 0.5 when validation loss plateaus for 3 consecutive epochs. The batch size is set to 32 for all experiments. The dataset is divided into 70% training, 10% validation, and 20% test sets.

To leverage both visual and textual

information for gambling promotion detection, this study explores six multimodal fusion strategies: concatenation, average fusion, weighted fusion, stacking fusion, projection fusion, and contrastive similarity fusion. These strategies span three fusion paradigms: early fusion, which combines embeddings before classification (concatenation); intermediate fusion, which transforms embeddings into a shared space before combining (projection, contrastive similarity); and late fusion, which combines predictions from separate unimodal classifiers (average, weighted, stacking).

The simplest fusion approach concatenates the image and text embeddings into a single unified representation. Given an image embedding $\mathbf{e}_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$ and a text embedding $\mathbf{e}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$, the fused representation is:

$$\mathbf{e}_{\text{fused}} = [\mathbf{e}_{\text{img}} || \mathbf{e}_{\text{txt}}] \in \mathbb{R}^{d_{\text{img}}+d_{\text{txt}}} \quad (1)$$

where $\|$ denotes concatenation. This fused vector is then passed through the MLP classifier. The concatenation approach preserves all information from both modalities but increases the input dimensionality.

Average fusion combines the prediction logits from separate unimodal classifiers through the arithmetic mean:

$$\text{logit}_{final} = \frac{1}{2}(f_{img}(\mathbf{e}_{img}) + f_{txt}(\mathbf{e}_{txt})) \quad (2)$$

where f_{img} and f_{txt} are independent MLP classifiers for each modality. This approach assumes equal contribution from both modalities without learnable weighting, serving as a baseline for more sophisticated fusion strategies.

Weighted fusion combines the prediction logits from separate unimodal classifiers through learned weights. Each modality has its own MLP classifier, and the final logit is computed as:

$$\text{logit}_{\text{final}} = w_1 \cdot f_{img}(\mathbf{e}_{img}) + w_2 \cdot f_{txt}(\mathbf{e}_{txt}) \quad (3)$$

where f_{img} and f_{txt} are independent MLP classifiers, and w_1, w_2 are learnable weight parameters normalized via softmax so that $w_1 + w_2 = 1$. The weights are initialized to equal contribution ($w_1 = w_2 = 0.5$) and optimized during training, allowing the model to learn the relative importance of each modality. This approach assumes that a linear combination of unimodal predictions is sufficient for the final decision.

Stacking fusion operates at the decision level rather than the feature level. Separate classifiers are trained for each modality independently, and their prediction probabilities are combined:

$$p_{\text{final}} = \beta \cdot p_{img} + (1 - \beta) \cdot p_{txt} \quad (4)$$

where p_{img} and p_{txt} are the predicted probabilities from the image-only and text-only classifiers, respectively, and β controls the modality contribution. Alternatively, a meta-classifier can be trained on the concatenated probability outputs. This approach allows each modality to be processed optimally by its own classifier while enabling flexible combination strategies.

Projection fusion maps both modality embeddings to a shared representation space using non-linear projections before concatenation. Each modality is transformed through a projection head consisting of a linear layer, ReLU activation, and dropout:

$$\mathrm{h}_{\text{img}} = \text{Dropout}(\text{ReLU}(\mathrm{W}_{\text{img}}\mathrm{e}_{\text{img}})) \quad (5)$$

$$\mathrm{h}_{\text{txt}} = \text{Dropout}(\text{ReLU}(\mathrm{W}_{\text{txt}}\mathrm{e}_{\text{txt}})) \quad (6)$$

where $W_{\text{img}} \in \mathbb{R}^{d \times d_{\text{img}}}$ and $W_{\text{txt}} \in \mathbb{R}^{d \times d_{\text{txt}}}$ project to a shared dimension $d = 512$. The projected representations are concatenated and passed to the classifier: $\mathbf{h}_{\text{fused}} = [\mathbf{h}_{\text{img}} || \mathbf{h}_{\text{txt}}]$. This approach allows each modality to learn task-specific transformations while maintaining the expressiveness of the original embeddings through non-linear projections.

Inspired by the contrastive learning paradigm of CLIP (Radford et al., 2021), this fusion strategy projects both modality embeddings into a shared semantic space and

measures their alignment through cosine similarity. Given image and text embeddings, linear projections are learned, followed by L2 normalization:

$$\mathbf{z}_{\text{img}} = \frac{W_{\text{img}}\mathbf{e}_{\text{img}}}{\|W_{\text{img}}\mathbf{e}_{\text{img}}\|_2}, \quad \mathbf{z}_{\text{txt}} = \frac{W_{\text{txt}}\mathbf{e}_{\text{txt}}}{\|W_{\text{txt}}\mathbf{e}_{\text{txt}}\|_2} \tag{7}$$

where $W_{\text{img}} \in \mathbb{R}^{d \times d_{\text{img}}}$ and $W_{\text{txt}} \in \mathbb{R}^{d \times d_{\text{txt}}}$ project to a shared dimension $d = 512$. The classification logit is computed as the scaled cosine similarity:

$$\text{logit} = \tau \cdot (\mathbf{z}_{\text{img}} \cdot \mathbf{z}_{\text{txt}}) \tag{8}$$

where τ is a learnable temperature parameter initialized to $1/0.07 \approx 14.3$. This approach captures the semantic alignment between visual and textual content, which is particularly effective for detecting gambling promotions where images and text often convey complementary promotional messages.

The performance of each image, text, and multimodal classifier was evaluated on the held-out test set using accuracy, precision, recall, and F1-score. These metrics were calculated from the confusion matrix for binary classification, where true positive (TP) denotes gambling promotion samples correctly classified as gambling promotion, true negative (TN) denotes non-promotion samples correctly classified as non-promotion, false positive (FP) denotes non-promotion samples incorrectly classified as gambling promotion, and false negative (FN) denotes gambling promotion samples incorrectly classified as non-promotion. The metric values reported in Tables 4, 5, and 6 were computed as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{11}$$

$$\text{F1-score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{12}$$

F1-score was used as the primary summary metric because this study focuses on detecting a target class of practical concern, namely online gambling promotions, where model performance should reflect both the reliability of positive predictions and the ability to identify actual promotion content. Precision indicates how many samples predicted as gambling promotions were correct, whereas recall indicates how many actual gambling promotion samples were successfully detected. Therefore, emphasizing the F1-score provides a balanced summary of these two complementary aspects. Nevertheless, accuracy, precision, and recall were also reported alongside the F1-score to provide a complete empirical evaluation of each model.

RESULTS AND DISCUSSION

Image Classification

Table 4 presents the image classification results. SigLIP 2 achieves the best F1-score of 93.52% and accuracy of 94.43%, followed by CLIP-ViT-L/14 at 92.74% and DINOv3-H+ at 89.74%. While CLIP obtains a slightly higher recall of 93.17% compared to SigLIP 2's 92.03%, SigLIP 2 compensates with a substantially higher precision of 95.06% versus 92.33%, resulting in a better overall F1-score. The two vision-language contrastive models outperform the self-supervised DINOv3, suggesting that pretraining with language supervision provides more discriminative visual features for this classification task. SigLIP 2's advantage over CLIP may be attributed to its higher input resolution and larger model capacity of 400M versus 307M parameters. Figure 7 shows the training loss curves for all three models.

Table 4. Image Classification Results on Test Set.

Model	Recall	Precision	F1-Score	Accuracy
DINOv3-H+	90.66%	88.84%	89.74%	90.95%
CLIP-ViT-L/14	93.17%	92.33%	92.74%	93.64%
SigLIP 2	92.03%	95.06%	93.52%	94.43%

Text Classification

Table 5 presents the text classification results. Multilingual E5-large achieves the best F1-score of 95.89% and accuracy of

96.42%, followed by Indonesian BERT (Cahya) at 94.49% and Qwen3-Embedding at 94.04%. Notably, despite being trained exclusively on Indonesian text, Cahya BERT achieves the highest precision of 97.34%, outperforming the much larger Qwen3-Embedding with 0.6B parameters. This suggests that language-specific pre-training provides strong representations, particularly for precision-sensitive tasks. However, Multilingual E5-large's contrastive pre-training on 1 billion multilingual text pairs yields the strongest overall semantic representations. Figure 8 shows the training loss curves for all three models.

Table 5. Text Classification Results on Test Set.

Model	Recall	Precision	F1-Score	Accuracy
Indonesian BERT	91.80%	97.34%	94.49%	95.33%
Qwen3-Embedding	93.39%	94.69%	94.04%	94.83%
Multilingual E5-large	95.67%	96.11%	95.89%	96.42%

Multimodal Classification

For multimodal classification, the best-performing unimodal encoders were used as the basis for all fusion experiments. SigLIP 2 was selected as the image encoder, while Multilingual E5-large was selected as the text encoder. Using the same pair of encoders across all experiments allows the comparison to focus on the fusion mechanism rather than differences in unimodal representation quality. Six fusion strategies were evaluated, namely concatenation, average, weighted, stacking, projection, and contrastive similarity fusion.

Table 6 summarizes the test-set performance of the six fusion strategies using recall, precision, F1-score, and accuracy. Contrastive similarity fusion achieved the strongest overall result, with a recall of 97.95%, F1-score of 97.18%, and accuracy of 97.51%. Projection fusion produced the highest precision of 98.82%, indicating that it was the most conservative strategy when assigning samples to the gambling promotion class. However, contrastive similarity fusion provided the best overall balance between identifying gambling promotion posts and maintaining high classification accuracy, making it the most effective multimodal strategy in this experiment.

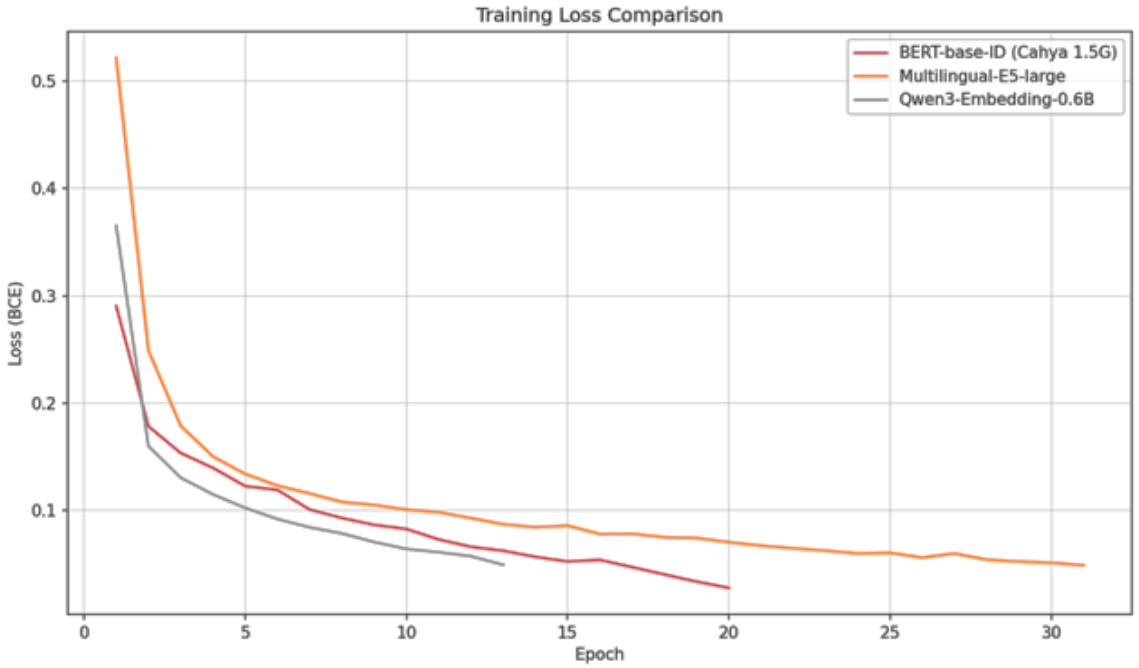


Figure 7. Training loss curves for image classification models: DINOv3-H+, CLIP-ViT-L/14, and SigLIP 2

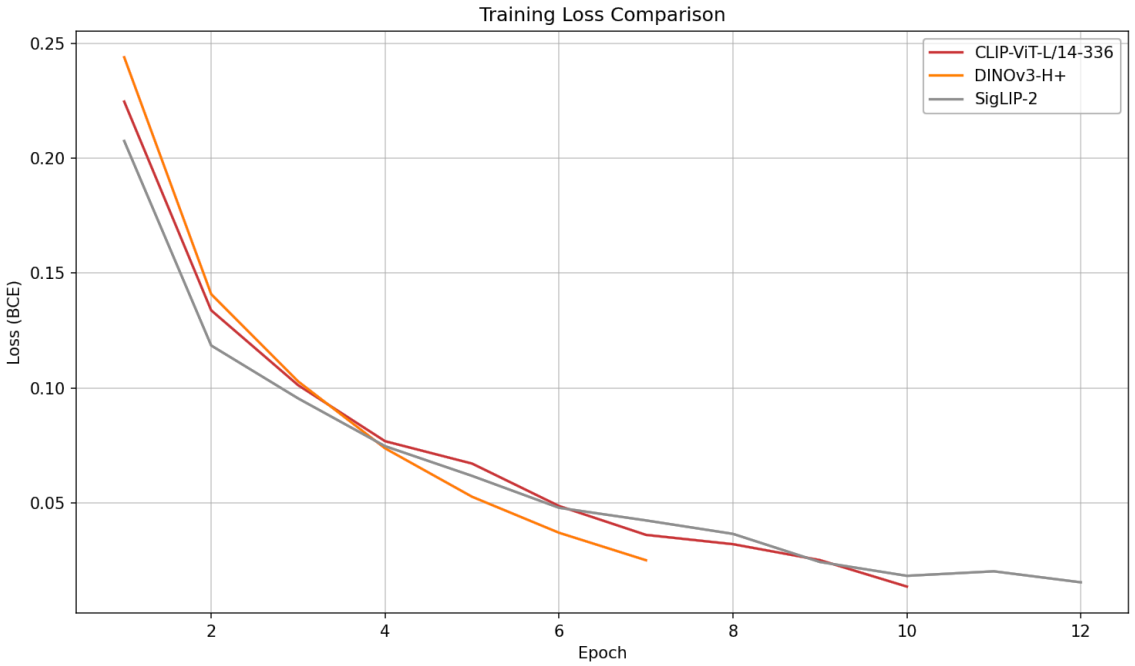


Figure 8. Training loss curves for text classification models: Indonesian BERT, Qwen3-Embedding, and E5-large.

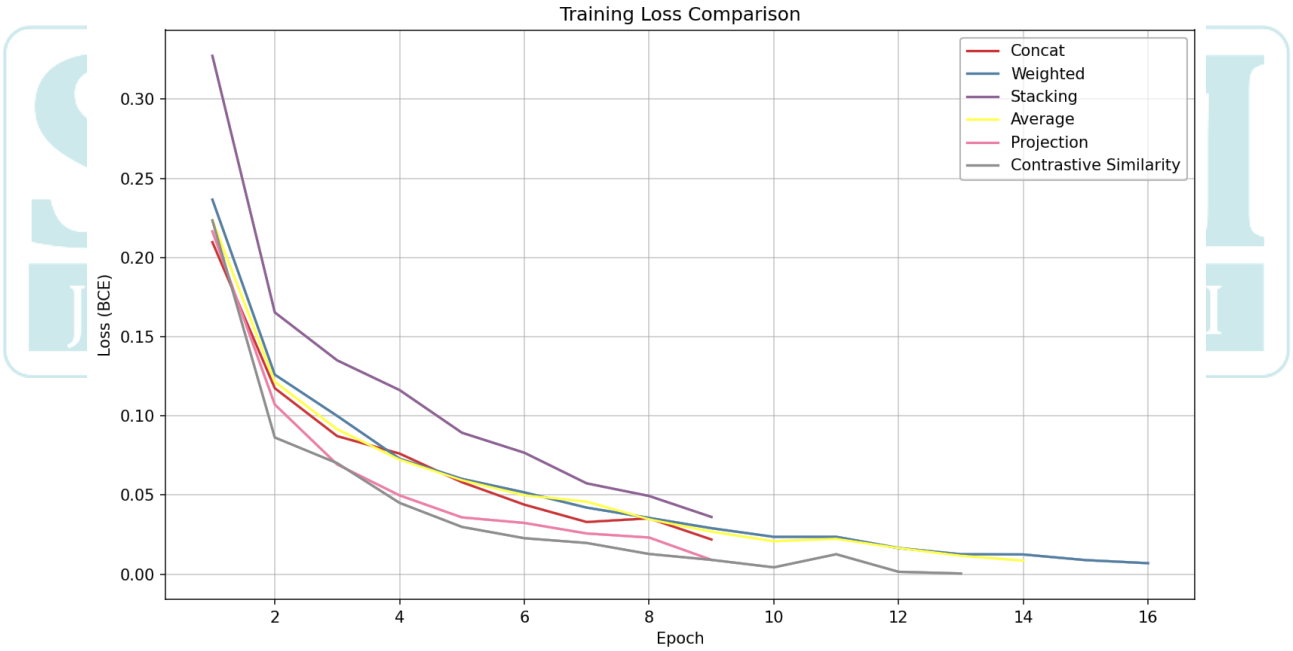


Figure 9. Training loss curves for six multimodal fusion strategies on the test set.

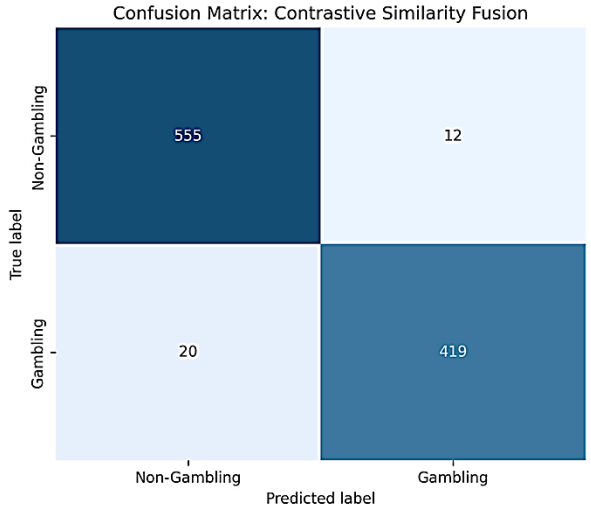


Figure 10. Confusion matrix for the Contrastive Similarity fusion model.

The ROC curve in Fig. 9 provides an additional diagnostic view of the best-performing Contrastive Similarity model across different decision thresholds, while the confusion matrix in Fig. 10 summarizes its final prediction distribution at the selected threshold. At this threshold, the model correctly classified 419 gambling promotion samples and 555 non-promotion samples, while producing 12 false positives and 20 false negatives. These results support the aggregate metrics in Table 6 by showing that most test samples were assigned to the correct class and that the remaining errors were limited relative to the total test set. Overall, the multimodal results indicate that image and text information provide complementary signals for detecting online gambling promotions, and that cross-modal semantic alignment is particularly effective for Indonesian social media posts where promotional intent is often conveyed through both captions and visual advertisement elements.

Table 6. Multimodal Classification Results on Test Set.

Fusion Strategy	Recall	Precision	F1-Score	Accuracy
Concatenation	95.44%	96.10%	95.77%	96.32%
Average	96.13%	97.91%	97.01%	97.42%
Weighted	96.13%	96.57%	96.35%	96.82%
Stacking	94.99%	93.71%	94.34%	95.03%
Projection	95.22%	98.82%	96.98%	97.42%
Contrastive Similarity	97.95%	96.41%	97.18%	97.51%

Ablation Study

Table 7 presents the ablation comparison between the best image-only, text-only, and multimodal configurations. The text-only Multilingual E5-large model outperformed the image-only SigLIP 2 model, indicating that captions provide stronger discriminative signals than visual content alone for this task. However, the multimodal Contrastive Similarity fusion achieved the highest F1-score and accuracy, showing that visual information still contributes complementary cues when it is aligned with textual representations. This result also suggests that performance gains do not come merely from adding another modality, but from using a fusion mechanism that can effectively model cross-modal semantic alignment.

Table 7. Ablation Study of Unimodal and Multimodal Configuration.

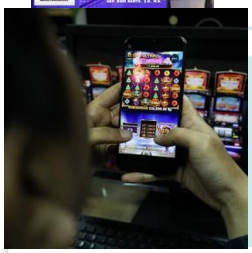
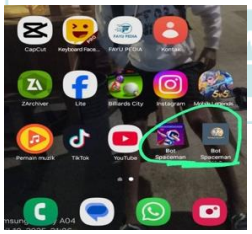
Fusion Strategy	Recall	Precision	F1-Score	Accuracy
Image-only (SigLIP 2)	95.44%	96.10%	95.77%	96.32%
Text-only (E5-large)	96.13%	97.91%	97.01%	97.42%
Multimodal (Constrastive Similarity)	96.13%	96.57%	96.35%	96.82%

Error Analysis

To complement the aggregate evaluation, we conducted a qualitative error analysis on the best-performing multimodal configuration, namely SigLIP 2 and E5-large with Contrastive Similarity fusion. Table 8 presents representative true positive, true negative, false positive, and false negative predictions from the test set. These cases should be interpreted alongside the multimodal classification results in Table 6, where the model achieved high recall, precision, F1-score, and accuracy. Therefore, the examples do not indicate a substantial degradation in performance, but instead

illustrate a small set of challenging cases that expose the remaining boundary conditions of the proposed model.

Table 8. Representative Prediction Outcomes From Constrative Similarity Fusion.

Image	Caption	Label	Pred
	SCATER BERKAH MAHJONG ROYALBET KING. ... bersama akun VIP.	1	1
	STOP JUDI ONLINE. 4 Bahaya Akibat Judi Online. ... @polresta_ bandarlampung.	0	0
	JUDI ONLINE HILANG. ... judi online dari platform digital.	0	1
	Predik bot spaceman. LIST HARGA: ... , 100K 1MINGGU, ALL SITUS.	1	0

The true positive example represents a straightforward gambling promotion, as both the image and caption contain explicit promotional cues, including Mahjong Ways, slot gacor, VIP account registration, jackpot-oriented wording, and an invitation to register or play. The true negative example, in contrast, contains gambling-related terms but is correctly classified as non-gambling because the overall context is an anti-gambling public service message that discusses financial loss, addiction, legal consequences, and mental health risks. The false positive case shows why lexical and visual cues alone can be insufficient: although the content is labeled non-gambling, the image depicts slot-machine activity, and the caption repeatedly mentions online gambling in a discussion or criticism context. Conversely, the false negative case shows the difficulty of detecting covert promotion when the image appears to be an ordinary smartphone home screen and the gambling signal is limited to small application icons, and domain-specific phrases such as Spaceman bot, price lists, and all-site availability.

Overall, the remaining errors are concentrated in semantically ambiguous cases rather than in clear promotional or clearly benign content. False positives tend to arise when non-promotional discussions, warnings, or news-like content contain strong gambling-related visual or textual cues. False negatives tend to occur when promotional intent is implicit, visually subtle, or expressed through domain-specific terms that may not be sufficiently salient in either modality. These findings suggest that the model is reliable for most cases, but borderline content may still benefit from threshold calibration, contextual post-processing, or human-in-the-loop review in practical moderation settings.

Policy and Government Implications

The empirical results suggest that the proposed Contrastive Similarity fusion model can support cybercrime mitigation by serving as an automated screening tool for online gambling promotion on social media. In practical deployment, the model can analyze newly collected public posts by jointly evaluating the image and caption, allowing it to detect promotional content that may not be identifiable from text alone. Such a system could be executed periodically, for example, at scheduled intervals, to monitor newly emerging public content and identify posts that require further inspection.

However, the model should not be used as a fully automated blocking mechanism. Instead, its output should be treated as a risk flag or priority score that helps moderators, platform operators, or relevant government agencies prioritize high-risk posts for review.

This human-in-the-loop workflow is important because the error analysis shows that some non-promotional posts, such as news, criticism, educational warnings, or anti-gambling campaigns, may still contain strong gambling-related terms or visuals and therefore require contextual judgment before enforcement action is taken. This approach can improve monitoring efficiency while reducing the risk of over-enforcement against legitimate informational or preventive content.

CONCLUSION

Conclusion

This study presents a multimodal deep learning approach for detecting online gambling promotions on Indonesian social media using a publicly available dataset of 5,028 labeled samples from Facebook, TikTok, and X. The empirical results show that text representations provide stronger detection signals than image representations, while multimodal fusion further improves classification by combining complementary visual and textual information. Among the evaluated approaches, Contrastive Similarity fusion between SigLIP 2 and Multilingual E5-large produced the strongest overall performance, with F1-score used as the main summary metric because it balances precision and recall in a binary detection task where both false positives and false negatives are important. These findings indicate that the proposed dataset and fusion-based detection system can support future empirical studies and may serve as a foundation for automated screening tools in real-world social media moderation pipelines.

Suggestions

This study also has several limitations that should be considered when interpreting the results. The dataset was collected using keyword-based sampling, which may introduce bias toward posts containing explicit gambling-related terms, and the dataset size remains limited relative to the diversity of online gambling promotion strategies. In addition, the data were obtained from Facebook, TikTok, and X, so model generalization to other platforms or future changes in platform-specific content patterns requires further validation. Future research should therefore evaluate larger and more diverse datasets, examine attention-based fusion mechanisms that dynamically weight visual and textual contributions, investigate newer text, image, and multimodal embedding models, and compare embedding-based fusion with end-to-end vision-language models such as Qwen3-VL (Bai et al., 2025; Yin et al., 2024) to determine whether more recent representations or multimodal reasoning paradigms provide additional empirical benefits for domain-specific gambling promotion detection.

BIBLIOGRAPHY

Alfiansyah, M. I., & Muzakir, A. (2025). Enhanced detection of Indonesian online gambling advertisements using multimodal ensemble deep learning. International *Journal of Artificial Intelligence and Science*, 2(2), 159–175. https://doi.org/10.63158/ijais.v2i2.49

Arifat, M., Putri, W. A., & Mufida, A. S. (2023). Penerapan metode Naive Bayes Classifier untuk klasifikasi indeks Pembangunan manusia di Provinsi Jawa Timur. *Jurnal Statistika Dan Komputasi*, 2(1), 31–43. https://doi.org/10.32665/statkom.v2i1.1661

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for Zero-Shot Cross-Lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597–610. https://doi.org/10.1162/tacl_a_00288

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., . . . Zhu, K. (2025). QWEN3-VL Technical Report. *ArXiv.org*. https://doi.org/10.48550/arxiv.2511.21631

Cahyo, D. D. N., Handayani, R., Lestari, V. B., & Febriani, S. (2025). Sentiment Analysis of Public Opinion on Online Gambling Through Social Media Using Convolutional Neural Network. *Journal of Information Technology and Engineering (JITE)*, 9(1), 99–115. https://doi.org/10.31289/jite.v9i1.15024

Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., & Xu, B. (2023). VLP: A Survey on vision-Language Pre-training. *Machine Intelligence Research*, 20(1), 38–56. https://doi.org/10.1007/s11633-022-1369-5

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., Elsheikh, A. H., Abualigah, L., & Al-qaness, M. A. A. (2021). Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics*, 10(11), 1332. https://doi.org/10.3390/electronics10111332

Fisher, M. L., Piper, T., Fitzpatrick, M., Mavi, S., Retzer, A., Bradbury-Jones, C., Montgomery, P., Melendez-Torres, G. J., Kirby, J., Chandan, J. S., & Bedford, K. (2025). Legal and regulatory responses to online gambling harms: a scoping review of evidence. *Harm Reduction Journal*, 22(1), 163. https://doi.org/10.1186/s12954-025-01292-y

Ginasputri, H. N. A., Saputri, A. D., Wahid, S. N., Ningrum, A. F., & Haris, M. A. (2025). Analisis sentimen ulasan pengguna BRIMO terhadap pembaruan fitur aplikasi menggunakan Naive Bayes dengan seleksi fitur Chi-Square. *Jurnal Statistika Dan Komputasi*, 4(2), 56–67. https://doi.org/10.32665/statkom.v4i2.5194

Gupta, S., & Kishan, B. (2025). A performance-driven hybrid text-image classification model for multimodal data. *Scientific Reports*, 15(1), 11598. https://doi.org/10.1038/s41598-025-95674-8

Ilham, M. L. (2025). Indonesian Financial Watchdog Finds $58.9 Million in Online Gambling Transactions Linked to State Social Aid Recipients. https://www.jakartadaily.id/market-finance/16215502616/indonesian-financial-watchdog-finds-589-million-in-online-gambling-transactions-linked-to-state-social-aid-recipients

Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. *International Journal of Advanced Computer Science and Applications*, 14(8). https://doi.org/10.14569/ijacsa.2023.0140813

Junaedi, J. (2024). England is the Largest Center for Online Gambling Activity in the World, Versus Indonesia is Exposed to Online Gambling Emergency Stage Five. *International Journal of Law, Crime and Justice*, 1(3), 100–114. https://doi.org/10.62951/ijlcj.v1i3.134

Kamdan, K., Anugrah, M. P., Almutaali, M. J., Ramdani, R., & Kharisma, I. L. (2025). Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments. *Engineering Proceedings*, 107(1), 66. https://doi.org/10.3390/engproc2025107066

Kongah, M. N. (2025). Promensisko 2025: Responding to the Threat of Online Gambling and Digital Crime Through Action. https://www.ppatk.go.id/siaran_pers/read/1474/promensisko-2025-menjawab-ancaman-judi-online-dan-kejahatan-digital-lewat-aksi-.html

Liem, J. A., Utomo, D. T. D., Ignasio, B., & Cenggoro, T. W. (2025). Deep Learning And Machine Learning For Indonesian Online Gambling Website Promotion Detection. *Procedia Computer Science*, 269, 979–992. https://doi.org/10.1016/j.procs.2025.09.040

Lin, S., Chen, Y., Chang, Y., Lo, S., & Chao, K. (2024). Text–image multimodal fusion model for enhanced fake news detection. *Science Progress*, 107(4), 368504241292685. https://doi.org/10.1177/00368504241292685

Lu, Y., & Yao, N. (2025). A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network. *Scientific Reports*, 15(1), 20544. https://doi.org/10.1038/s41598-025-05702-w

Maldini, A. S., Saputra, W. S. J., & Prasetya, D. A. (2025). Multimodal Detection of Covert Online Gambling Advertisements Using Faster R-CNN and Tr-OCR. *Bit-Tech*, 8(1), 953–963. https://doi.org/10.32877/bt.v8i1.2769

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

Muggleton, N., Parpart, P., Newall, P., Leake, D., Gathergood, J., & Stewart, N. (2021). The association between gambling and financial, social and health outcomes in big financial data. *Nature Human Behaviour*, 5(3), 319–326. https://doi.org/10.1038/s41562-020-01045-w

Muzakir, A., Ependi, U., & Suyanto. (2025). A deep learning and AutoML-Based multimodal text extraction framework for detecting online gambling advertisements in Indonesian social media. *International Journal of Safety and Security Engineering*, 15(9). https://doi.org/10.18280/ijsse.150908

Perdana, R. B., Ardin, 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 and Applications*, 15(8). https://doi.org/10.14569/ijacsa.2024.0150893

Prabhu, R., & Seethalakshmi, V. (2025). A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, 15(1), 13020. https://doi.org/10.1038/s41598-025-94069-z

Pratama, H. M., Wijayakusuma, I. L., & Widiastuti, R. S. (2025). Comparison of online gambling promotion detection performance using DistilBERT and DeBERTA models. *Journal of Applied Informatics and Computing*, 9(6), 3716–3725. https://doi.org/10.30871/jaic.v9i6.11293

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models from Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, 139, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Rahman, M. R. (2025). Indonesia pressed to intensify fight against online gambling networks. https://en.antaranews.com/news/390781/indonesia-pressed-to-intensify-fight-against-online-gambling-networks

Santika, E. F. (2025). Gambling-related Divorces in Indonesia Soar 83% in 2024. https://databoks.katadata.co.id/en/demographics/statistics/689ad79cdce8e/gambling-related-divorces-in-indonesia-soar-83-in-2024

Siméoni, O., Vo, H., V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., . . . Bojanowski, P. (2025). DINOv3. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.10104

Situmorang, G. F., Manurung, J. D., Lubis, R., Sikana, N., & Lionggo, K. (2026). An intelligent system for detecting online gambling promotion on social media based on BERT and adaptive browser extension. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, 14(3). https://doi.org/10.23887/janapati.v14i3.103511

Stankevičius, L., & Lukoševičius, M. (2024). Extracting Sentence Embeddings from Pretrained Transformer Models. *Applied Sciences*, 14(19), 8887. https://doi.org/10.3390/app14198887

Sutriawan, Rustad, S., Shidik, G. F., & Pujiono. (2025). Performance Evaluation of text embedding models for ambiguity classification in Indonesian news Corpus: A comparative study of TF-IDF, Word2VEC, FastText BERT, and GPT. *Ingénierie Des Systèmes D Information*, 30(6), 1469–1482. https://doi.org/10.18280/isi.300606

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., & Zhai, X. (2025). SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *ArXiv.org*. https://doi.org/10.48550/arxiv.2502.1

4786

Uelwer, T., Robine, J., Wagner, S. S., Höftmann, M., Upschulte, E., Konietzny, S., Behrendt, M., & Harmeling, S. (2025). A survey on self-supervised methods for visual representation learning. *Machine Learning*, 114(4). https://doi.org/10.1007/s10994-024-06708-7

Ukhova, D., Marionneau, V., Nikkinen, J., & Wardle, H. (2023). Public health approaches to gambling: a global review of legislative trends. *The Lancet Public Health*, 9(1), e57–e67. https://doi.org/10.1016/s2468-2667(23)00221-9

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: a Technical report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.05672

Widiyanto, A., Prameswari, M., & Latief, M. A. (2025). Gambling comments detection on YouTube: A comparative study of Tree-Based Boosting, LSTM and GRU models. *JUTI Jurnal Ilmiah Teknologi Informasi*, 144–160. https://doi.org/10.12962/j24068535.v23i2.a1305

Wirawan, C. (2022). Indonesian BERT base model (uncased) – 1.5G. https://huggingface.co/cahya/bert-base-indonesian-1.5G

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, 11(12), nwae403. https://doi.org/10.1093/nsr/nwae403

Zhang, X., Li, G., Tu, S., Yi, J., & Zhang, Y. (2025). A gambling website identification framework based on fusion-assisted multimodal large language model. *Journal of Computational Methods in Sciences and Engineering*. https://doi.org/10.1177/14727978251364464

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). QWeN3 Embedding: advancing text embedding and reranking through foundation models. *ArXiv.org*. https://doi.org/10.48550/arxiv.2506.05176