Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

# Multimodal Emotion Classification of Indonesian Memes on Platform X Using IndoBERT and YOLOv11

**Pungkas Subarkah[1], Esti Widianti[2], Agus Pramono[3]**

[1,2,3]Informatics Department, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia

**Received:**
February 6 2026
**Revised:**
June 15, 2026
**Accepted:**
July 26 2026
**Published:**
August 22, 2026

Corresponding Author:
**Author Name*:**
Pungkas Subarkah
**Email*:**
subarkah@amikompurwokerto.ac.id

DOI:
10.63158/journalisi.v8i4.1742



**Abstract.** The widespread use of social media, particularly Platform X, has increased the popularity of memes as a multimodal communication medium that combines textual and visual elements to express emotions, opinions, and social reactions. This study proposes a multimodal emotion classification approach for Indonesian-language memes by integrating IndoBERT for textual analysis and YOLOv11 operating in image classification mode for visual analysis through a weighted late fusion strategy. An initial dataset of 2,810 Indonesian-language memes was collected through web scraping. After removing corrupted or unreadable images, 2,547 valid samples remained. Each meme was manually annotated into one of six emotion categories—Happiness, Disgust, Anger, Sadness, Fear, and Surprise—based on the annotators' judgment of the dominant emotion conveyed by the combination of text and image, following Ekman's Basic Emotion framework. The dataset was divided using a stratified 80:10:10 split into 2,037 training, 255 validation, and 255 testing samples. The validation set was used to determine the optimal fusion weight, while the held-out test set was reserved exclusively for final evaluation. The best-performing weighted late fusion model ($\alpha$ = 0.6) achieved 72.55% accuracy, 73.03% macro precision, 72.76% macro recall, and 72.66% macro F1-score on the test set. Within the constructed dataset, the proposed multimodal approach outperformed the evaluated IndoBERT-only and YOLOv11-only baselines, indicating that combining textual and visual information can improve emotion classification performance for Indonesian-language meme content.

**Keywords**: Multimodal Emotion Classification, Indonesian Memes, IndoBERT, YOLOv11, Weighted Late Fusion

## 1. INTRODUCTION

The rapid advancement of digital technology has fundamentally transformed the way people communicate, share information, and express emotions through social media platforms. Beyond functioning as communication channels, social media has evolved into an interactive digital environment where users continuously generate multimodal content consisting of text, images, videos, and other visual elements [1]. According to the Digital 2025 Global Overview Report, the number of global social media users reached approximately 5.24 billion in 2025, demonstrating the increasingly significant role of social media in everyday communication and public interaction [2]. As a result, social media has become an important source of large-scale user-generated data for computational research, particularly in studies focusing on public opinion, sentiment, emotion, and online behavioral analysis.

Among the various social media Platform X (formerly Twitter) occupies a distinctive position as a public communication space where users actively express opinions, emotions, humor, and reactions to current events in real time through short textual posts accompanied by images, videos, and other multimedia content [3]. The 2026 Digital Report recorded approximately 22.9 million X users in Indonesia at the beginning of 2026, making Indonesia one of the largest markets for this platform in Southeast Asia [2]. This substantial user base indicates that Platform X remains a valuable source of multimodal social media data for computational analysis because it contains diverse forms of public expression that reflect emotional, cultural, and social interactions within Indonesian online communities [4].

One of the most prominent forms of multimodal communication on Platform X is the widespread use of internet memes [5]. Rather than serving merely as humorous images, memes have evolved into a distinctive communication medium that combines textual and visual information to convey emotions, opinions, satire, social criticism, and cultural references in a concise yet expressive manner [6]. Previous studies have shown that memes play an increasingly important role in representing collective emotional responses to various events, including political discussions, public health issues, natural disasters, and other social phenomena [7]. In the Indonesian context, memes circulated during the 2024 general election demonstrated how visual and textual elements can

simultaneously influence public opinion, reinforce political narratives, and reflect cultural identity within online communities [8]. These findings suggest that memes provide a valuable resource for understanding emotional expression in digital communication.

The multimodal nature of memes makes emotion classification considerably more challenging than conventional text-based sentiment analysis [9]. Emotional meaning is often constructed through the interaction between textual content and visual elements rather than through either modality alone. Meme captions frequently contain slang, abbreviations, sarcasm, irony, emojis, and culturally specific expressions, while the accompanying images provide contextual cues such as facial expressions, gestures, symbolic objects, and visual composition [10]. Consequently, unimodal approaches that analyze only text or only images often fail to capture the complete emotional context of a meme. Previous studies, including the Memotion benchmark, have consistently demonstrated that multimodal learning strategies outperform single-modality approaches because they exploit complementary information from both textual and visual representations [5], [7]. These findings highlight the importance of developing multimodal approaches for emotion classification, particularly for Indonesian-language memes, whose linguistic and cultural characteristics remain substantially underexplored [11].

A closer look at prior work shows a clear split by modality. Text-only approaches, such as sentiment and sarcasm detection with IndoBERT, capture linguistic cues but miss visual signals that often carry the punchline of a meme [11]. Image-only approaches, including YOLO-based facial-expression recognition, capture visual cues but miss the wordplay, slang, and cultural references embedded in the caption [12]. Multimodal approaches, as shown by Sharma et al. and by Bhapkar's Memotion 2.0 project, consistently outperform single-modality models precisely because meme emotion is jointly encoded in text and image [13], [14]. This study builds on that multimodal line of work but targets Indonesian-language memes specifically, a setting largely absent from the existing literature, which is dominated by English-language datasets [4], [1].

Previous studies on meme emotion analysis have generally followed three research directions: text-based, image-based, and multimodal approaches. Text-based methods employ natural language processing models to analyze captions or embedded text

extracted from memes, whereas image-based methods rely solely on visual information [15]. Although both approaches have shown promising results in emotion and sentiment classification tasks, each has inherent limitations because memes encode meaning through the interaction of textual and visual elements rather than through a single modality [16]. Recent multimodal studies have therefore demonstrated that integrating both modalities produces more comprehensive representations and generally achieves higher classification performance than unimodal approaches [17].

Despite these advances, current research remains dominated by English-language datasets, such as the Memotion benchmark, while studies focusing on Indonesian-language memes are still scarce. Existing Indonesian research has primarily concentrated on sentiment analysis of textual social media posts or image classification independently, with relatively few studies addressing multimodal emotion classification for Indonesian memes [18]. Furthermore, there is currently no widely recognized benchmark dataset specifically designed for Indonesian-language multimodal meme emotion classification [5]. Consequently, it remains unclear whether multimodal architectures developed and evaluated on English-language datasets can effectively capture the linguistic characteristics, cultural context, humor, and emotional expressions commonly found in Indonesian memes [19]. This gap highlights the need for research specifically designed for Indonesian-language multimodal content.

To address this challenge, this study employs IndoBERT as the text encoder [20]. IndoBERT is a transformer-based language model pre-trained exclusively on large-scale Indonesian corpora, enabling it to better capture Indonesian vocabulary, grammar, semantic relationships, and contextual language patterns than multilingual models such as mBERT or XLM-R [19]. Previous studies have demonstrated that IndoBERT performs effectively in various Indonesian natural language processing tasks, including sentiment analysis, sarcasm detection, and emotion classification, particularly when processing informal language commonly found on social media [21]. Because meme captions frequently contain slang, abbreviations, spelling variations, and culturally specific expressions, IndoBERT provides a suitable textual representation for emotion classification in Indonesian-language memes [11].

For visual analysis, this study adopts YOLOv11 operating in classification mode. Although the YOLO family is traditionally associated with object detection, its classification variant utilizes the same efficient convolutional backbone to learn discriminative visual representations from entire images without performing object localization [12]. YOLOv11 provides competitive image classification performance while maintaining high computational efficiency and fast inference speed. Previous studies have also reported that YOLO-based architectures can effectively capture visual cues associated with emotional expressions, including facial expressions, scene composition, and contextual visual patterns relevant to emotion recognition tasks [22]. These characteristics make YOLOv11 an appropriate visual feature extractor for multimodal meme emotion classification.

The combination of IndoBERT and YOLOv11 is expected to exploit complementary information from textual and visual modalities through a multimodal learning framework that combines complementary information from heterogeneous modalities [1]. While textual information provides semantic understanding of captions and embedded messages, visual information contributes contextual cues that may not be explicitly expressed in text alone [23]. By integrating both modalities through a weighted late fusion strategy, the proposed framework aims to improve the robustness and accuracy of emotion classification compared with single-modality approaches [24], [25]. Nevertheless, this study does not assume that the proposed framework represents all Indonesian memes, as the dataset was constructed from publicly available content collected from Platform X during a specific collection period. Therefore, the reported findings should be interpreted within the scope of the constructed dataset rather than as a comprehensive representation of Indonesian online culture.

Finally, considering the increasing volume of multimodal content on social media, an effective emotion classification framework has practical implications beyond academic research. Such a framework can support applications including social media monitoring, public opinion analysis, digital culture research, and decision-support systems that require automatic interpretation of emotional expressions in large-scale multimodal online content [26]. These practical needs further motivate the development of an Indonesian-language multimodal emotion classification framework capable of integrating textual and visual information within a unified learning architecture [4].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Based on the limitations identified in previous studies, there remains a clear need for a multimodal emotion classification framework specifically designed for Indonesian-language memes [4]. Existing studies have predominantly focused on English-language datasets, such as Memotion and HarMeme, while Indonesian-language multimodal meme datasets remain very limited [13]. Furthermore, many previous approaches either analyze textual and visual information independently or evaluate multimodal architectures on datasets with different linguistic and cultural characteristics, limiting their applicability to Indonesian social media content [1]. Since memes often convey emotions through sarcasm, irony, and culturally contextual expressions, an effective emotion classification system should jointly interpret textual and visual information rather than relying on a single modality.

To address this research gap, this study proposes a multimodal emotion classification framework that integrates IndoBERT for textual feature extraction and YOLOv11 operating in classification mode for visual feature extraction. The prediction probabilities generated by both models are subsequently combined using a weighted late fusion strategy, which has been widely adopted in multimodal learning to exploit complementary information from different modalities while maintaining relatively simple model integration [27]. Unlike previous studies that primarily focused on English-language meme datasets, the proposed framework is specifically designed for Indonesian-language memes collected from Platform X, allowing the evaluation of multimodal learning under Indonesian linguistic and cultural characteristics [28].

The novelty of this research lies in four main aspects. First, this study constructs one of the largest manually annotated Indonesian-language multimodal meme datasets based on Ekman's six basic emotions [29]. Second, it develops a multimodal emotion classification framework that integrates IndoBERT and YOLOv11 using a weighted late fusion strategy. Third, it provides an empirical comparison between text-only, image-only, and multimodal approaches using the same dataset. Finally, this study presents an initial analysis of emotion distribution in Indonesian-language memes while discussing methodological challenges associated with multimodal emotion classification in culturally contextual social media environments [28].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The primary objective of this study is to investigate whether combining textual and visual information through weighted late fusion can improve emotion classification performance compared with the evaluated unimodal models [30]. Specifically, this research evaluates the effectiveness of integrating IndoBERT and YOLOv11 for classifying Indonesian-language memes into six emotion categories based on Ekman's Basic Emotion framework [29]. Model performance is evaluated using Accuracy, Macro Precision, Macro Recall, and Macro F1-score, which are widely recognized as standard evaluation metrics for multi-class classification problems.

Because this study utilizes publicly available user-generated meme content collected from Platform X, ethical considerations were incorporated throughout the research process. Only meme images and their embedded textual content were analyzed, while personally identifiable information—including usernames, profile information, account metadata, and direct links to the original posts—was excluded from the dataset to minimize potential privacy risks [31]. The collected data were used solely for academic research purposes and are presented only in aggregate form, following commonly accepted ethical practices in computational social media research [32]. Furthermore, this study does not claim that the constructed dataset fully represents Indonesian meme culture; rather, the reported findings should be interpreted within the scope of the collected dataset and the defined data collection period.

Overall, this study is expected to contribute both theoretically and practically to the advancement of multimodal emotion classification for Indonesian-language social media content. From a theoretical perspective, it provides empirical evidence regarding the effectiveness of integrating transformer-based language models and deep visual classifiers for multimodal emotion recognition [33]. From a practical perspective, the proposed framework has potential applications in social media monitoring, public opinion analysis, digital culture studies, and intelligent decision-support systems requiring automatic interpretation of emotional expressions in multimodal online content[34].

## 2. METHODS

This study proposes a multimodal framework for classifying emotions in Indonesian-language memes collected from Platform X by integrating textual and visual information using IndoBERT, YOLOv11 in image classification mode, and a weighted late fusion strategy. The overall methodology was designed to systematically construct the dataset, develop independent unimodal classifiers, and evaluate the effectiveness of multimodal fusion for emotion classification.

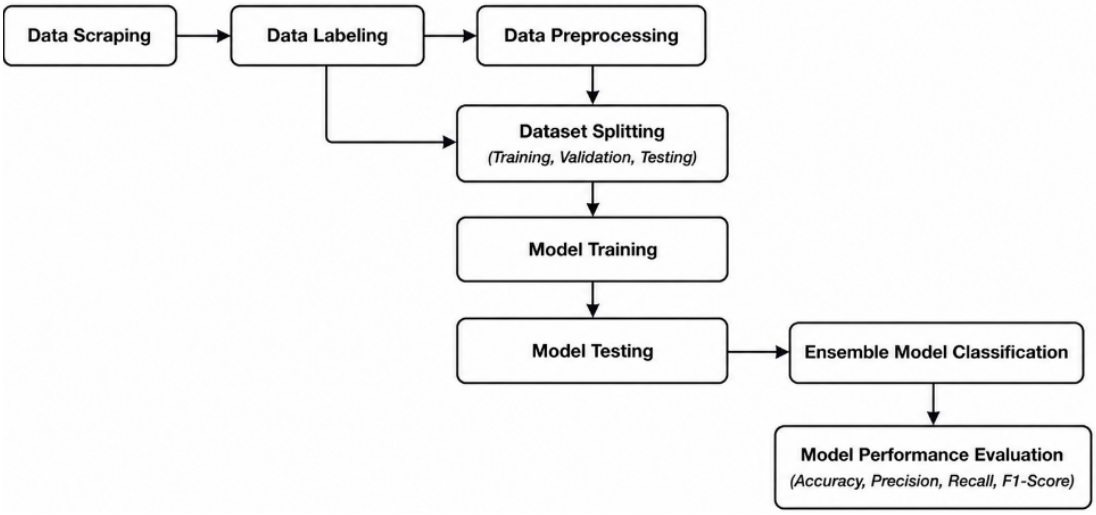


**Figure 1.** Research Methodology Workflow

As illustrated in Figure 1, the proposed framework consists of several sequential stages. The study begins with dataset construction through web scraping, followed by manual emotion annotation based on Ekman's six basic emotions. The collected data are then preprocessed separately for the textual and visual modalities before being partitioned into training, validation, and testing subsets. Subsequently, the textual and visual classifiers are trained independently, after which their prediction probabilities are combined using a weighted late fusion approach. Finally, the proposed framework is evaluated using standard classification metrics on a held-out test dataset. The detailed implementation of each stage is described in Sections 2.1–2.8.

### 2.1. Research Data

The research data used in this study consist of Indonesian-language meme content collected from Platform X through a web scraping process. Each collected sample

contains two complementary modalities, namely a meme image and the textual content embedded within the image, which serve as the primary sources of information for multimodal emotion classification. Data collection was conducted between February and March 2026, resulting in an initial dataset of 2,810 meme samples.

To facilitate the emotion classification task, the collected memes were organized into six emotion categories based on Ekman's Basic Emotion theory, namely Happiness, Disgust, Anger, Sadness, Fear, and Surprise [29]. These emotion categories were selected because they provide a well-established and widely adopted framework for emotion recognition, while remaining suitable for manual annotation of multimodal meme content. The distribution of the collected dataset across the six target emotion categories is presented in Table 1.

**Table 1.** Emotion Dataset Distribution

| No | Emotion Class | Number of Data | Percentage |
|---|---|---|---|
| 1. | Happiness | 500 | 17.79% |
| 2. | Disgust | 500 | 17.79% |
| 3. | Anger | 500 | 17.79% |
| 4. | Sadness | 500 | 17.79% |
| 5. | Surprise | 500 | 17.79% |
| 6. | Fear | 310 | 11.03% |

Table 1 shows that five emotion categories (Happiness, Disgust, Anger, Sadness, and Surprise) each contain 500 collected samples, whereas the Fear category contains 310 samples. The collected dataset subsequently underwent data quality control, annotation, and preprocessing before being used for model development. The detailed procedures for these stages are described in the following sections.

### 2.2. Data Collection and Annotation

Data collection was performed through a web scraping process to retrieve Indonesian-language meme content from Platform X [35]. Meme candidates were collected using Indonesian keywords and hashtags representing the six target emotion categories (Happiness, Disgust, Anger, Sadness, Fear, and Surprise) within the January–December 2025 collection period. Each retrieved post consisted of a meme image containing

embedded textual content. The scraping process was implemented using a Python-based crawler that automatically collected post URLs, downloaded meme images, and organized the retrieved data according to the corresponding emotion category.

Following data collection, Optical Character Recognition (OCR) was performed using the Tesseract OCR engine through the pytesseract library with the ind+eng language configuration to extract textual content from each meme image [36]. The extracted text subsequently underwent quality control through text normalization, language verification, image inspection, and perceptual hash (pHash) deduplication to remove unreadable, non-Indonesian, corrupted, and duplicate samples. After these procedures, the dataset was reduced from 2,810 collected samples to 2,547 valid samples for annotation and subsequent model development.

The filtered dataset was manually annotated using Label Studio based on Ekman's six basic emotions. Annotation was performed by a single annotator, who assigned one dominant emotion label (Happiness, Disgust, Anger, Sadness, Fear, or Surprise) to each meme according to the combined interpretation of its textual and visual content. The same image-level emotion label was subsequently used as the ground-truth label for both the IndoBERT text classifier and the YOLOv11 image classifier. During annotation, text and object regions were additionally marked to facilitate manual inspection; however, these regions were not used as training targets because YOLOv11 was implemented in image classification mode rather than object detection mode. Since only one annotator was involved, inter-annotator agreement was not calculated, which is acknowledged as a limitation of this study [37].

### 2.3. Text Preprocessing

Text preprocessing was performed to prepare the extracted meme text for input into the IndoBERT model. After the OCR and quality-control procedures described in Section 2.2, the extracted text underwent a series of preprocessing steps to improve data consistency while preserving the semantic information required for emotion classification. The preprocessing pipeline began with text normalization, including the removal of unnecessary whitespace and non-informative characters. Subsequently, text cleaning was applied to eliminate URLs, user mentions, punctuation marks, hashtags, emojis, and other irrelevant symbols that were not required for the classification task.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

All text was then converted to lowercase to maintain a consistent representation across the dataset. Examples of the text cleaning results are presented in Table 2.

**Table 2.** Text Cleaning Results

| No | Original Text | Cleaned Text |
|---|---|---|
| 1. | *mentertawakan kebodohan diri sendiri | mentertawakan kebodohan diri sendiri |
| 2. | salting* | salting |
| 3. | Makasii | makasii |
| 4. | HAHAHAHA!!! | hahahaha |
| ... | ... | ... |
| 2547. | Sumpah Demi Apa??!! | sumpah demi apa |

After text cleaning, the processed text was tokenized using the IndoBERT tokenizer to generate input_ids, attention_mask, and token_type_ids, which served as inputs to the IndoBERT model during training and inference [38]. In addition, the emotion labels were converted into numerical class indices using the LabelEncoder implementation from the scikit-learn library to ensure consistent label representation during model training and evaluation.

### 2.4. Image Preprocessing

Image preprocessing was performed to prepare the visual data for training the YOLOv11 image classification model. Following the quality-control procedures described in Section 2.2, all images were inspected to ensure that they were readable, correctly labeled, and free from corruption before being used for model development. Subsequently, each image was resized to 224 × 224 pixels, corresponding to the default input resolution of the YOLOv11 classification model [39]. This resizing process ensured a consistent input dimension for both the training and inference stages while maintaining compatibility with the model architecture.

After preprocessing, the images were organized into separate directories according to their emotion labels (Happiness, Disgust, Anger, Sadness, Fear, and Surprise) following the dataset structure required by the YOLOv11 classification framework. This directory organization enabled the model to automatically associate each image with its

corresponding class label during training and evaluation, thereby ensuring consistency between the visual data and the emotion labels used throughout the multimodal classification framework.

### 2.5. Dataset Splitting

Following the preprocessing stage, the final dataset consisting of 2,547 annotated meme samples was partitioned into three subsets for model development and evaluation. To ensure that each emotion category was proportionally represented across all subsets, a stratified splitting strategy was employed [40]. The dataset was divided into 80% training, 10% validation, and 10% testing, thereby maintaining a balanced distribution of the six emotion classes while minimizing potential sampling bias. The detailed distribution of samples for each emotion category after the dataset partitioning process is presented in Table 3.

**Table 3.** Dataset Splitting Results

| Emotion Label | Total | Training | Validation | Testing |
|---|---|---|---|---|
| Happiness | 478 | 382 | 48 | 48 |
| Disgust | 459 | 367 | 46 | 46 |
| Anger | 482 | 386 | 48 | 48 |
| Sadness | 403 | 322 | 40 | 41 |
| Fear | 295 | 236 | 30 | 29 |
| Surprise | 430 | 344 | 43 | 43 |
| Total | 2547 | 2037 | 255 | 255 |

The training set was used to train both the IndoBERT and YOLOv11 models by optimizing their respective model parameters. The validation set was utilized throughout the training process to monitor model performance, perform hyperparameter tuning, and identify the best-performing model while reducing the risk of overfitting. Finally, the testing set was reserved exclusively for the final performance evaluation and was not involved in either model training or hyperparameter optimization. This separation ensured that the reported evaluation metrics accurately reflected the generalization capability of the proposed multimodal emotion classification framework when applied to previously unseen data.

### 2.6. Model Training

The proposed multimodal framework was trained using two independent deep learning models, namely IndoBERT for textual emotion classification and YOLOv11 for visual emotion classification. Both models were trained separately using the training dataset described in Section 2.5, while the validation dataset was employed to monitor model performance and select the optimal model before the fusion stage. For textual classification, the pretrained indobenchmark/indobert-base-p1 model proposed by Wilie et al. [11]. The maximum input sequence length was set to 128 tokens, with a batch size of 16, learning rate of $2 \times 10^{-5}$, weight decay of 0.01, and 7 training epochs. Model optimization was performed using the AdamW optimizer. To reduce overfitting, an Early Stopping strategy with a patience value of 2 epochs was applied, and the model with the highest validation Macro F1-score was selected as the final text classification model. For visual classification, the YOLOv11 classification model was trained using an input image size of 224 × 224 pixels, a batch size of 16, 22 training epochs, and an Early Stopping patience of 10 epochs. The validation dataset was used to monitor training performance and determine the best-performing model checkpoint. After both models had been trained independently, the best-performing IndoBERT and YOLOv11 models were used to generate prediction probabilities, which were subsequently combined through the weighted late fusion approach described in Section 2.7.

### 2.7. Weighted Late Fusion

A weighted late fusion strategy was employed to combine the prediction probabilities generated independently by the IndoBERT and YOLOv11 models, which is a widely used approach in multimodal classification for integrating complementary information from different modalities [41]. For the IndoBERT branch, the class probabilities were obtained by applying the Softmax function to the output logits, a standard probability estimation technique for deep learning classifiers [24]. Meanwhile, the YOLOv11 branch produced class confidence scores through the image classification model. The final probability for each emotion class was calculated using Equation (1):

$$P_{combined} = \alpha \times P_{YOLO} + (1 - \alpha) \times P_{IndoBERT} \quad (1)$$

where $P_{YOLO}$ and $P_{IndoBERT}$ represent the class probability vectors generated by the YOLOv11 and IndoBERT models, respectively, while $\alpha$\alpha$\alpha$ denotes the weighting coefficient assigned to the visual model.

The optimal value of α was determined by evaluating values ranging from 0.1 to 0.9 using the validation dataset. For each candidate value, the weighted fusion output was evaluated based on validation accuracy and macro F1-score, and the value that produced the best overall performance was selected. The evaluation results indicated that α = 0.6 achieved the highest validation performance, assigning 60% weight to the YOLOv11 model and 40% weight to the IndoBERT model. This optimal weighting was subsequently fixed and applied to the held-out test dataset for the final evaluation without further adjustment, thereby preventing information leakage from the test set during model selection.

### 2.8. Evaluation Metrics

The performance of the proposed multimodal emotion classification framework was evaluated using the held-out test dataset consisting of 255 samples that were not involved in model training, hyperparameter tuning, or alpha selection. Model performance was assessed using four standard classification metrics, namely accuracy, precision, recall, and F1-score, which are widely adopted for evaluating multi-class classification models. To provide a fair evaluation across all emotion categories, macro-averaged precision, recall, and F1-score were calculated by assigning equal weight to each class regardless of its sample size. In addition to the overall evaluation metrics, confusion matrices and per-class precision, recall, and F1-score were analyzed to assess the classification performance of each emotion category individually. These evaluation results were used to compare the performance of the IndoBERT, YOLOv11, and the proposed weighted late fusion model, thereby providing a comprehensive assessment of the effectiveness of the proposed multimodal framework.

## 3. RESULTS AND DISCUSSION

### 3.1. YOLOv11 Training

The YOLOv11 image classification model was trained for 22 epochs using a pretrained classification checkpoint and evaluated on the held-out test dataset. The model achieved an accuracy of 68.24%, macro precision of 69.38%, macro recall of 69.32%, and macro F1-score of 68.28%, indicating that visual features alone were able to capture a substantial amount of emotional information contained in Indonesian meme images.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

**Table** 4. Per-Class Classification Performance of the YOLOv11 Model on the Test Dataset

| Classification Repoort | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **Happines** | 0,82 | 0,56 | 0,67 | 48 |
| **Disgust** | 0,60 | 0,78 | 0,68 | 46 |
| **Anger** | 0,68 | 0,54 | 0,60 | 48 |
| **Sadness** | 0,82 | 0,80 | 0,81 | 41 |
| **Fear** | 0,56 | 0,79 | 0,66 | 29 |
| **Surprise** | 0,67 | 0,67 | 0,67 | 43 |
| | | | | |
| **Accuracy** | - | - | 0,68 | 255 |
| **Macro avg** | 0,69 | 0,69 | 0,68 | 255 |
| **Weighted avg** | 0,70 | 0,68 | 0,68 | 255 |

Table 4 presents the per-class performance of the YOLOv11 model on the held-out test dataset. The model achieved the highest F1-score for the Sadness class (0.81), indicating that this emotion was represented by more distinctive visual patterns such as facial expressions and scene composition. In contrast, the Anger class obtained the lowest F1-score (0.60), suggesting greater visual similarity with other negative emotions, particularly Fear and Disgust, resulting in more frequent misclassification. The remaining emotion classes achieved F1-scores ranging from 0.66 to 0.68, demonstrating relatively balanced performance across most categories. Overall, the model achieved a macro F1-score of 68.28% and an accuracy of 68.23%, indicating that visual information alone provides a strong baseline for emotion classification in Indonesian meme images, although it remains insufficient to fully capture the semantic information conveyed through textual content.

The relatively higher performance for the Sadness category indicates that this emotion is represented by more distinctive visual patterns, such as facial expressions, crying scenes, or gloomy visual compositions that can be effectively learned by the image classifier. Conversely, Anger shares several visual characteristics with other negative emotions, including Fear and Disgust, causing greater ambiguity during classification. Since the model relies solely on image features, contextual information embedded in meme text is not considered, which may reduce its ability to distinguish emotions with similar visual appearances but different semantic meanings. Overall, these findings

demonstrate that the visual modality provides a strong baseline for meme emotion recognition. However, the remaining classification errors indicate that image information alone cannot fully represent the multimodal characteristics of memes, thereby motivating the integration of textual features through the proposed multimodal fusion approach.

### 3.2. IndoBERT Training Results

The IndoBERT model was evaluated using the held-out test dataset consisting of 255 Indonesian meme samples. The evaluation employed accuracy, macro precision, macro recall, and macro F1-score to measure the effectiveness of text-based emotion classification using textual information extracted through Optical Character Recognition (OCR). The overall evaluation results are presented in Table 5.

**Table** 5. Overall Performance of the IndoBERT Model

| Metrix | Value |
|---|---|
| Accuracy | 0,5333 |
| Precision Macro | 0,6648 |
| Recall Macro | 0,5202 |
| F1-Score Macro | 0,5307 |

The IndoBERT model achieved an overall accuracy of 53.33% with a macro F1-score of 53.07%, indicating that textual information alone was less effective than visual information for emotion classification in Indonesian memes. Compared with the YOLOv11 image classifier, which achieved a macro F1-score of 68.28%, the text-based model produced substantially lower performance across most evaluation metrics. This result suggests that meme text alone often does not provide sufficient semantic information to accurately represent the intended emotional expression. The detailed per-class performance is presented in Table 6.

Table 6 shows that the Sadness class achieved the highest F1-score (0.78), indicating that textual expressions associated with sadness were relatively more distinctive and easier for IndoBERT to recognize. In contrast, the Fear class obtained the lowest F1-score (0.41) despite having relatively high precision (0.80) because its recall was only 0.28. This indicates that although predictions of the Fear class were generally correct when

produced, the model failed to identify many Fear samples, resulting in numerous false negatives. The remaining emotion classes achieved F1-scores between 0.44 and 0.57, reflecting moderate performance and considerable variation across emotion categories. One important factor contributing to the lower performance of IndoBERT is the quality of the textual input obtained through OCR. Meme images frequently contain stylized fonts, low-resolution text, overlapping visual elements, or decorative typography that can introduce recognition errors during OCR. These errors may produce incomplete or incorrect textual content, thereby reducing the semantic information available for the language model. In addition, many Indonesian memes rely on sarcasm, irony, colloquial expressions, and visual context, making the emotional meaning difficult to infer from text alone. Overall, these findings indicate that textual information extracted from meme images provides complementary rather than sufficient information for emotion classification. The relatively lower performance of IndoBERT compared with YOLOv11 motivates the integration of textual and visual representations through the proposed weighted late fusion strategy, which is evaluated in the following section.

**Table 6.** Classification Report of IndoBERT on the Test Dataset

| Classification Repoort | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Happines | 0,75 | 0,31 | 0,44 | 48 |
| Disgust | 0,33 | 0,83 | 0,47 | 46 |
| Anger | 0,55 | 0,60 | 0,57 | 48 |
| Sadness | 0,88 | 0,71 | 0,78 | 41 |
| Fear | 0,80 | 0,28 | 0,41 | 29 |
| Surprise | 0,68 | 0,40 | 0,50 | 43 |
| Accuracy | - | - | 0,53 | 255 |
| Macro avg | 0,66 | 0,52 | 0,53 | 255 |
| Weighted avg | 0,65 | 0,53 | 0,53 | 255 |

### 3.3. Alpha Optimization for Weighted Late Fusion

After determining the optimal fusion weight ($\alpha = 0.6$) using the validation dataset, the proposed weighted late fusion model was evaluated on the held-out test dataset consisting of 255 meme samples. The evaluation was conducted without further parameter adjustment to assess the model's ability to generalize to previously unseen data. Performance was measured using accuracy, macro precision, macro recall, and

macro F1-score, as these metrics provide a comprehensive assessment of multi-class classification performance. As presented in Table 7, the fusion model showed a steady improvement as the contribution of the visual modality increased from $\alpha = 0.1$ to $\alpha = 0.6$. The best validation performance was obtained at $\alpha = 0.6$, achieving a validation accuracy of 72.55% and a macro F1-score of 72.66%. When the weight assigned to YOLOv11 was increased beyond $\alpha = 0.6$, both accuracy and macro F1-score gradually decreased, indicating that excessive reliance on visual information reduced the complementary contribution of the textual modality.

**Table 7.** Performance at Different Alpha Values (Validation Set)

| Alpha (α) | YOLOv11 Weight | IndoBERT Weight | Accuracy | F1-Score |
|---|---|---|---|---|
| 0.1 | 10% | 90% | 55.29% | 54.81% |
| 0.2 | 20% | 80% | 58.04% | 57.62% |
| 0.3 | 30% | 70% | 61.57% | 61.18% |
| 0.4 | 40% | 60% | 65.10% | 64.75% |
| 0.5 | 50% | 50% | 68.63% | 68.31% |
| 0.6 | 60% | 40% | 72.55% | 72.66% |
| 0.7 | 70% | 30% | 71.37% | 71.12% |
| 0.8 | 80% | 20% | 70.20% | 69.88% |
| 0.9 | 90% | 10% | 69.41% | 69.05% |

The selected $\alpha = 0.6$ assigns 60% of the final prediction weight to YOLOv11 and 40% to IndoBERT. This weighting is consistent with the stronger standalone performance of YOLOv11 observed in Section 3.1 while still preserving the semantic information extracted from meme text by IndoBERT. The results indicate that visual information contributes more substantially to emotion recognition in Indonesian memes, whereas textual information provides complementary contextual cues that improve the overall classification performance. After the optimal alpha value had been selected using the validation dataset, the same configuration ($\alpha = 0.6$) was applied without further adjustment during the final evaluation on the held-out test dataset. Although the validation performance at $\alpha = 0.6$ is numerically identical to the final test performance in this experiment, the two evaluations were performed on different datasets and at different stages of the experimental pipeline. The validation dataset was used solely for alpha optimization, whereas the test dataset remained untouched until the final

evaluation, thereby preventing information leakage between model optimization and performance assessment.

### 3.4. Final Evaluation of the Weighted Late Fusion Model

After determining the optimal fusion weight ($\alpha = 0.6$) using the validation dataset, the proposed weighted late fusion model was evaluated on the held-out test dataset consisting of 255 meme samples. The evaluation was conducted without further parameter adjustment to assess the model's ability to generalize to previously unseen data. Performance was measured using accuracy, macro precision, macro recall, and macro F1-score, as these metrics provide a comprehensive assessment of multi-class classification performance.

**Table 8.** Performance of late fusion model

| Metrix | Value |
|---|---|
| Best Alpha | 0,6 |
| Accuracy | 0,7255 |
| Precision Macro | 0,7303 |
| Recall Macro | 0,7276 |
| F1-Score Macro | 0,7266 |

As presented in Table 8, the proposed weighted late fusion model achieved an accuracy of 72.55%, with a macro precision of 73.03%, macro recall of 72.76%, and macro F1-score of 72.66%. Compared with the standalone YOLOv11 model, which achieved an accuracy of 68.23%, the proposed multimodal approach improved the overall classification accuracy by 4.32 percentage points. The improvement over the IndoBERT model was even more substantial, increasing from 53.33% to 72.55%. These results indicate that combining visual and textual information enables the model to capture complementary emotional cues that cannot be fully represented by either modality alone.

The superior performance of the weighted late fusion model can be attributed to the complementary characteristics of the two modalities. YOLOv11 effectively extracts visual features such as facial expressions, objects, colors, and meme layouts, whereas IndoBERT captures semantic information contained in the textual content. By combining prediction probabilities from both models with the optimized weighting configuration (60% YOLOv11

and 40% IndoBERT), the proposed approach benefits from the strengths of each modality while reducing their individual limitations. This finding is consistent with the motivation of multimodal emotion classification, where emotional meaning is conveyed through both visual and textual information.

Table 9 shows that the proposed weighted late fusion model achieved relatively balanced performance across the six emotion categories. The Sadness class obtained the highest classification performance with an F1-score of 0.83, followed by the Disgust class with an F1-score of 0.80. These results suggest that memes expressing sadness and disgust tend to contain more distinctive combinations of visual and textual characteristics, enabling the multimodal model to classify them more consistently.

**Table 9.** Per-Class Performance of the Combined Model (α = 0.6)

| Emotion Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Happiness | 0.74 | 0.73 | 0.73 | 48 |
| Disgust | 0.80 | 0.80 | 0.80 | 46 |
| Anger | 0.68 | 0.65 | 0.66 | 48 |
| Sadness | 0.84 | 0.83 | 0.83 | 41 |
| Fear | 0.65 | 0.69 | 0.67 | 29 |
| Surprise | 0.67 | 0.63 | 0.65 | 43 |
| Macro Average | 0.73 | 0.73 | 0.73 | 255 |

In contrast, the Anger, Fear, and Surprise classes achieved relatively lower F1-scores. This performance difference indicates that these emotions share more overlapping visual expressions and semantic contexts, making them more difficult to distinguish. For example, memes expressing anger and disgust frequently exhibit similar facial expressions or negative wording, while surprise and fear may share comparable visual reactions or contextual cues. Consequently, the model experiences greater difficulty in separating these emotion categories despite the integration of visual and textual information.

### 3.5. Confusion Matrix Analysis

To further examine the classification behavior of the proposed weighted late fusion model, a confusion matrix was generated using the predictions on the held-out test

dataset [42]. The confusion matrix provides detailed information regarding correctly classified samples as well as misclassification patterns among the six emotion categories.

Table 10 shows that most predictions are concentrated along the main diagonal, indicating that the majority of meme samples were correctly classified into their corresponding emotion categories. The Disgust class achieved the highest number of correct predictions with 37 of 46 samples (80.4%), followed by Sadness with 34 of 41 samples (82.9%), and Happiness with 33 of 48 samples (68.8%). These results suggest that memes expressing disgust and sadness contain relatively distinctive visual and textual characteristics that enable the multimodal model to classify them more consistently.

**Table 10.** Confusion Matrix

| Actual \ Predicted | Happy | Disgust | Angry | Sad | Fear | Surprise |
|---|---|---|---|---|---|---|
| **Happy** | 33 | 2 | 8 | 4 | 0 | 1 |
| **Disgust** | 0 | 37 | 2 | 1 | 3 | 3 |
| **Angry** | 4 | 7 | 31 | 2 | 0 | 4 |
| **Sad** | 0 | 2 | 2 | 34 | 2 | 1 |
| **Fear** | 2 | 1 | 2 | 0 | 21 | 3 |
| **Surprise** | 1 | 7 | 2 | 0 | 4 | 29 |

Misclassification primarily occurred among emotion categories that share similar visual expressions or semantic meanings. The Anger class was most frequently misclassified as Disgust (7 samples; 14.6%) and Happiness (4 samples; 8.3%). Likewise, the Surprise class was often predicted as Disgust (7 samples; 16.3%) and Fear (4 samples; 9.3%). These confusion patterns indicate that several emotion categories exhibit overlapping visual features and contextual language, making them more difficult to distinguish even when visual and textual modalities are combined. Overall, the confusion matrix confirms that the proposed weighted late fusion model achieved balanced classification performance across the six emotion categories while maintaining relatively low levels of misclassification. Nevertheless, confusion remained for emotion pairs with closely related semantic and visual characteristics, indicating that further improvements may be achieved through larger and more balanced datasets or more advanced multimodal fusion strategies.

### 3.6. Emotion Distribution in Classification Results

After evaluating the proposed weighted late fusion model on the held-out test dataset, the model was further applied to the entire preprocessed corpus consisting of 2,547 meme samples to examine the overall distribution of predicted emotions. Since the weighted late fusion approach requires valid probability outputs from both the YOLOv11 and IndoBERT branches, 281 samples (11.0%) could not be assigned a fused prediction because no usable Indonesian text remained after the OCR extraction and text preprocessing stages. Consequently, these samples were excluded from this analysis, resulting in 2,266 samples with valid multimodal predictions. The classification results reveal the following distribution across emotion categories.

As shown in Table 11, Happiness was the most frequently predicted emotion, accounting for 19.0% of all classified memes, followed by Disgust (18.5%) and Surprise (17.8%). In contrast, Fear was the least frequently predicted emotion, representing 11.9% of the classified samples. This distribution is generally consistent with the composition of the constructed dataset, where fear-related memes were less prevalent than the other emotion categories. Overall, the predicted emotion distribution indicates that the proposed model did not exhibit a strong bias toward a single emotion class, with predictions remaining relatively balanced across the six emotion categories.

**Table 11.** Emotion Distribution of Classification Results

| Emotion Category | Number of Samples | Percentage |
|---|---|---|
| Happiness | 431 | 19.0% |
| Disgust | 419 | 18.5% |
| Anger | 366 | 16.2% |
| Sadness | 377 | 16.6% |
| Surprise | 403 | 17.8% |
| Fear | 270 | 11.9% |
| Total | 2,266 | 100% |

### 3.7. Discussion

The experimental results demonstrate that the proposed multimodal approach, which combines YOLOv11 and IndoBERT using weighted late fusion, achieves better emotion classification performance than either unimodal model. The combined model improved

the overall accuracy from 68.23% for YOLOv11 and 53.33% for IndoBERT to 72.55%, indicating that visual and textual information provide complementary cues for emotion recognition. This finding is consistent with previous studies reporting that multimodal approaches generally outperform unimodal methods in sentiment and emotion analysis by integrating heterogeneous sources of information [31].

The superior performance of YOLOv11 can be attributed to its improved feature extraction capability, including the C3k2 block architecture and the SPPF module, which enable the model to capture discriminative visual features from meme images more effectively [14]. In contrast, IndoBERT achieved lower performance because meme captions are often short, sarcastic, or highly dependent on the accompanying image for their emotional meaning. Furthermore, OCR extraction errors caused by stylized fonts, watermarks, or low-contrast text introduce additional noise into the textual input, reducing the effectiveness of text-only classification.

Overall, the results demonstrate that weighted late fusion successfully exploits the complementary strengths of the visual and textual modalities. Nevertheless, the reported performance reflects agreement with the manually annotated emotion labels rather than the actual emotional intentions of meme creators or audiences. Therefore, the findings should be interpreted within the context of the constructed dataset and should not be generalized beyond the annotated Indonesian meme corpus used in this study.

Although the proposed weighted late fusion approach achieved improved performance compared with the unimodal YOLOv11 and IndoBERT models, several limitations should be acknowledged. First, the dataset is moderately imbalanced, particularly for the Fear category, which contains fewer samples than the other emotion classes and may have contributed to the lower classification performance for this category. Second, the quality of OCR-extracted text varies considerably due to stylized fonts, watermarks, low-resolution images, and decorative text commonly found in memes. These factors introduce noise into the textual input and reduce the effectiveness of the IndoBERT branch. Third, the manually annotated emotion labels represent the dominant emotion perceived by human annotators. Therefore, the reported performance reflects agreement with the annotated labels rather than the actual emotional intention of meme creators or audiences.

Another limitation concerns the multimodal fusion procedure itself. Because the weighted late fusion model requires valid probability outputs from both the visual and textual branches, 281 samples (11.0% of the 2,547 preprocessed samples) could not be assigned a fused prediction after OCR extraction and text preprocessing produced no usable Indonesian text. Consequently, these samples were excluded from the emotion distribution analysis, reducing the effective sample size from 2,547 to 2,266.

Future studies should investigate probability calibration techniques, such as temperature scaling, before multimodal fusion [43]. This includes expanding the dataset with a larger and more balanced collection of Indonesian-language memes, investigating fallback strategies for samples without usable textual information, and exploring adaptive or end-to-end multimodal fusion methods. In addition, future studies should compare the proposed approach with stronger visual and textual baseline models, such as ResNet, EfficientNet, Vision Transformer (ViT), mBERT, and XLM-R, and evaluate the effect of probability calibration and repeated experimental trials to provide a more comprehensive assessment of multimodal emotion classification performance.

## 4. CONCLUSION

This study proposed a multimodal emotion classification framework for Indonesian memes by combining YOLOv11 for image classification and IndoBERT for text classification using a weighted late fusion strategy. The proposed approach achieved better performance than the evaluated unimodal YOLOv11-only and IndoBERT-only models on the constructed dataset, indicating that integrating visual and textual information can improve emotion classification for Indonesian memes. However, these findings are limited to the manually annotated dataset and the held-out test set used in this study, and should not be interpreted as evidence of general superiority over other multimodal approaches or broader Indonesian meme datasets. Although the results are promising, the improvement over the evaluated YOLOv11 baseline remains modest and requires further validation through statistical testing, repeated experiments, and evaluation on external datasets. Future work should investigate stronger visual and textual baseline models, probability calibration before multimodal fusion, inter-annotator agreement analysis, fallback strategies for samples without usable textual information,

and larger, more balanced datasets, particularly for the Fear category, further to improve the robustness and generalizability of the proposed framework.

## REFERENCES

[1] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Comput. Surv.*, vol. 55, no. 13s, Art. no. 270, Dec. 2023, doi: 10.1145/3586075.

[2] "Digital 2025: Global overview report," DataReportal—Global Digital Insights, 2025. [Online]. Accessed: Jul. 24, 2026.

[3] M. Ihsan and R. S. Adnan, "Media sosial Twitter sebagai ruang publik virtual (Studi kasus penolakan Omnibus Law)," *Syntax Lit.: J. Ilm. Indones.*, vol. 7, no. 3, pp. 3254–3267, Mar. 2022, doi: 10.36418/syntax-literate.v7i3.6612.

[4] U. Singh, K. Abhishek, and H. K. Azad, "A survey of cutting-edge multimodal sentiment analysis," *ACM Comput. Surv.*, vol. 56, no. 9, Art. no. 227, pp. 1–38, 2024, doi: 10.1145/3652149.

[5] C. Sharma *et al.*, "SemEval-2020 Task 8: Memotion analysis—The visuo-lingual metaphor!," in *Proc. 14th Int. Workshop Semantic Evaluation (SemEval)*, 2020, pp. 759–773, doi: 10.18653/v1/2020.semeval-1.99.

[6] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 4439–4455, doi: 10.18653/v1/2021.findings-emnlp.379.

[7] D. Kiela *et al.*, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 2611–2624, 2020.

[8] H. Uswatun and A. Ahmadi, "Meme politik di TikTok: Perspektif pragmatika humor sebagai kritik sosial digital," *J. Pendid. Bhs. Sastra Indones.*, vol. 2, no. 2, pp. 72–89, Jun. 2026, doi: 10.47134/jpbsi.v2i2.2625.

[9] A. Pandey and D. K. Vishwakarma, "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," *Appl. Soft Comput.*, vol. 152, Art. no. 111206, Feb. 2024, doi: 10.1016/j.asoc.2023.111206.

[10] T. Modi, E. Shah, S. Shah, J. Kanakia, and M. Tiwari, "Meme classification and offensive content detection using multimodal approach," in *Proc. 2024 OITS Int. Conf. Inf. Technol. (OCIT)*, 2024, pp. 635–640, doi: 10.1109/OCIT65031.2024.00116.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[11] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics and 10th Int. Joint Conf. Natural Lang. Process. (AACL-IJCNLP)*, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.

[12] Ultralytics, "Ultralytics YOLO11," Ultralytics Docs. 2026, [Online]. Available: https://docs.ultralytics.com/models/yolo11/. Accessed: Jul. 22, 2026.

[13] J. H. Chowdhury and S. Ramanna, "MMLTC: A novel tolerance-based clustering framework for multimodal sentiment and harmful meme classification in multilingual settings," *Comput. Intell.*, vol. 42, no. 2, Art. no. e70219, Mar. 2026, doi: 10.1111/coin.70219.

[14] N. P. Bhapkar, "Memotion 2.0—Sentiment analysis and emotion classification of memes," M.Sc. research project, Data Analytics, National College of Ireland, Dublin, Ireland, 2022.

[15] S. Patel, N. Shroff, and H. Shah, "Multimodal sentiment analysis using deep learning: A review," *Commun. Comput. Inf. Sci.*, vol. 2038, pp. 13–29, 2024, doi: 10.1007/978-3-031-59097-9_2.

[16] S. K. Baberwal, N. A. Shelke, and K. Anwar, "Systematic review of recent advances in multimodal sentiment analysis," *Discov. Comput.*, vol. 28, no. 1, Art. no. 270, Nov. 2025, doi: 10.1007/s10791-025-09727-7.

[17] M. Dhotay, M. Dharrao, S. Deokate, A. Bongale, and D. Dharrao, "Multimodal sentiment analysis: Emerging innovations, core challenges, and future directions," *Discov. Artif. Intell.*, vol. 6, no. 1, Art. no. 433, Mar. 2026, doi: 10.1007/s44163-026-01141-2.

[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguistics (COLING)*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[19] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[20] C. Shaw, P. M. LaCasse, and L. E. Champagne, "Exploring emotion classification of Indonesian tweets using large scale transfer learning via IndoBERT," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, Art. no. 22, Mar. 2025, doi: 10.1007/s13278-025-01439-6.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[21] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN," *ILKOM J. Ilm.*, vol. 14, no. 3, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022, doi: 10.1109/TAFFC.2020.2981446.

[23] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[24] A. Kumar, K. Sharma, and A. Sharma, "MEmoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries," *Image Vis. Comput.*, vol. 123, Art. no. 104483, Jul. 2022, doi: 10.1016/j.imavis.2022.104483.

[25] A. Nandi, F. Xhafa, L. Subirats, and S. Fort, "Reward-penalty weighted ensemble for emotion state classification from multi-modal data streams," *Int. J. Neural Syst.*, vol. 32, no. 12, Art. no. 2250049, Dec. 2022, doi: 10.1142/S0129065722500496.

[26] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D., "Multimodal sentimental analysis for social media applications: A comprehensive review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 11, no. 5, Art. no. e1415, Sep. 2021, doi: 10.1002/WIDM.1415.

[27] K. Zhao, M. Zheng, Q. Li, and J. Liu, "Multimodal sentiment analysis—A comprehensive survey from a fusion methods perspective," *IEEE Access*, vol. 13, pp. 64556–64583, 2025, doi: 10.1109/ACCESS.2025.3554665.

[28] R. K. Routhu and U. Baruah, "Sentiment analysis on memes: A review," *Expert Syst.*, vol. 42, no. 11, Art. no. e70133, Nov. 2025, doi: 10.1111/EXSY.70133.

[29] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, nos. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.

[30] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, Jul. 2023, doi: 10.1016/j.inffus.2023.02.028.

[31] M. L. Williams, P. Burnap, and L. Sloan, "Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation," *Sociology*, vol. 51, no. 6, pp. 1149–1168, Dec. 2017, doi: 10.1177/0038038517708140.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[32] A. N. Markham, K. Tiidenberg, and A. Herman, "Ethics as methods: Doing ethics in the era of big data research—Introduction," *Soc. Media Soc.*, vol. 4, no. 3, Jul. 2018, doi: 10.1177/2056305118784502.

[33] S. Hazmoune and F. Bougamouza, "Using transformers for multimodal emotion recognition: Taxonomies and state-of-the-art review," *Eng. Appl. Artif. Intell.*, vol. 133, Art. no. 108339, Jul. 2024, doi: 10.1016/j.engappai.2024.108339.

[34] Y. Dui and H. Hu, "Social media public opinion detection using multimodal natural language processing and attention mechanisms," *IET Inf. Secur.*, vol. 2024, no. 1, Art. no. 8880804, 2024, doi: 10.1049/2024/8880804.

[35] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jul. 2016, doi: 10.1145/2818717.

[36] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, Curitiba, Brazil, 2007, pp. 629–633, doi: 10.1109/ICDAR.2007.4376991.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[38] Y. Jiao, J. Li, J. Wu, D. Hong, R. Gupta, and J. Shang, "SeNsER: Learning cross-building sensor metadata tagger," in *Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 950–960, doi: 10.18653/v1/2020.findings-emnlp.85.

[39] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, 2019, pp. 6105–6114.

[40] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, QC, Canada, 1995, pp. 1137–1143.

[41] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimed. Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010, doi: 10.1007/s00530-010-0182-0.

[42] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[43] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 1321–1330.