Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal emotion recognition from images and text using deep learning and explainable AI

Fatima Uroosa
Instituto Politécnico Nacional
Muhammad Tayyab Zamir
Instituto Politécnico Nacional
Francesco Perono Cacciafoco
Xi'an Jiaotong-Liverpool University
Asim Abbas
University of Birmingham
Rolando Quintero Téllez
quintero@cic.ipn.mx

Instituto Politécnico Nacional
Grigori Sidorov
Instituto Politécnico Nacional

Article

Keywords: Multimodal emotion recognition, ITMER framework, Bidirectional Encoder Representations from Transformers (BERT), Explainable AI, Image– text emotion analysis

Posted Date: May 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9453940/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal emotion recognition from images and text using deep learning and explainable AI

Anonymous Author(s)

Abstract

Emotions play an important role in human life. A person's emotional state can be identified through emotional expression, whether they are sad, happy, or angry. Emotions can be expressed in various ways, including text, facial expressions, and audio signals. Additionally, emotion recognition plays a crucial role in affective computing and human-computer interaction, addressing the complex challenges of interpreting human emotions across diverse data sources. Although significant research has been conducted in emotion recognition, existing studies have focused on video-based or speech-based multimodal systems, with limited attention given to the efficient integration of textual and image data. Furthermore, most prior studies lack systematic unimodal–multimodal comparison and do not address model interpretability in the image–text emotion recognition framework. To address this gap, we propose ITMER (Image–Text Multimodal Emotion Recognition), a unified framework that integrates textual and visual modalities for emotion classification. The framework is evaluated under three settings: unimodal text-based emotion recognition using pre-trained transformer encoder models, unimodal image-based emotion recognition using deep convolutional neural network models, and multimodal emotion recognition through image–text fusion. The proposed study offers insight into the contribution of each modality and the efficacy of multimodal approaches by systematically comparing these configurations. Experimental results demonstrate that multimodal image–text learning consistently outperforms unimodal baselines, achieving up to 92% accuracy and a macro F1-score of 92%, compared to 83.81% for the best image-only and 71% for the best text-only model, highlighting the benefit of integrating complementary visual and linguistic information. To further improve transparency, explainable AI techniques, including Grad-CAM for visual models and LIME for text models, are applied post-hoc to provide qualitative insight into model decisions, validating that predictions are grounded in meaningful emotional cues rather than spurious patterns.

Keywords: Multimodal emotion recognition; ITMER framework; Bidirectional Encoder Representations from Transformers (BERT); Explainable AI; Image–text emotion analysis

1 Introduction

Emotions are complex, psychological, and physiological states that reflect humans' state of mind, and they can be influenced by various factors, such as Behavior and personality (Matsumoto and Hwang, 2016). Similarly, emotions are a fundamental part of human life, influencing life choices as well as psychological and physical health.(Kement et al., 2024)(Fei et al., 2020). Accordingly, emotion recognition has emerged as a multidisciplinary field that integrates psychology, computer science, and artificial intelligence. Moreover, humans express emotions through various channels: auditory, visual, and textual. Accurately interpreting these expressions is important for advancing human-computer interaction, mental health diagnosis, and security systems. From another perspective, emotions are essential regulators of human experience and behavior, influencing factors such as decision-making, Communication, and overall well-being, covering states like sadness, happiness, and anger (Feidakis et al., 2011).

The goal of affective computing focuses on training machines to recognize and respond to human emotions, thereby supporting more natural and effective human technology interactions.(Spezialetti et al., 2020). Emotion expression is a complex process, and it highlights the importance of multifaceted emotion recognition. Emotions are multidimensional and are usually expressed through different channels such as the voice tone, facial expression, and choice of words. Such intermodal emotional expression poses serious challenges to the computational recognition systems, which need to be integrated to examine and interpret complex emotional states from diverse data sources (Trinh and Dao, 2022). Consequently, over the last few years, multimodal emotion recognition (MER) has gained increased attention as it aims to overcome the limitations of single-channel approaches and use several sources of datasets.

The literature has demonstrated that utilizing multiple modalities (facial expressions, speech, and physiological cues significantly enhances emotion recognition performance. (Zhang et al., 2023). Recent advances in multimodal large language models (MLLMs) have further enabled emotion understanding to move beyond traditional discriminative approaches towards a more generative framework (Liang et al., 2024). However, existing MER systems still face notable limitations in fine-grained emotion understanding, particularly when applied to diverse, real-world data Lian et al. (2024). Likewise, studies on multimodal emotion recognition have focused on complex, video-centric systems that utilize time-series features extracted from multiple modalities, such as video and audio transcripts. For instance, (Lian et al., 2025) proposes AffectGPT, which focuses on complex emotion understanding with MLLMs applied to videos; while (Wang et al., 2022) focuses on specialized approaches integrating physiological data such as EEG alongside speech signals.

Similarly, (Kumar et al., 2025) combines video media with text, using complex ensembles (e.g., EfficientNetV2-S and GPT-2 LoRA) primarily designed for achieving high-accuracy classification in dynamic environments. However, these approaches involve high computational complexity and are not designed for the efficient integration of static image and textual data, which are the

dominant modalities in social media contexts. In particular, there is limited work that systematically combines transformer-based text encoders with deep visual models in a dual-encoder framework specifically designed for social media emotion recognition. To address this gap, this work investigates emotion recognition using both unimodal and multimodal approaches based on images and textual data. Large language models are applied independently to textual and visual modalities to establish strong unimodal baselines. A multimodal fusion framework is then constructed to integrate complementary emotional cues from both modalities. Through systematic comparison of unimodal and multimodal configurations, this study provides insights into the individual contribution of each modality and the effectiveness of multimodal fusion for social media–based emotion recognition.

The emphasis on combining effectiveness, explainability, and responsibility, and the focus on balancing efficacy and interpretability promotes the ethical application of AI. Understanding emotional context is essential in delicate fields like customer service and mental health monitoring, where the suggested framework is well-positioned for significant applications, and understanding emotional context is crucial. To enhance this work further, we consider describing how XAI approaches are integrated into an image–text emotion recognition framework to offer comprehensible insights on model forecasts in order to further improve this work. Moreover, deep learning approaches are currently constrained by the "black-box" problem of deep learning models. These models involve complex layers that operate as decision makers but are not visible to the onlookers, and operate in the background, making transparency a little difficult (Sahin et al., 2025).

Thus, the lack of transparency makes it difficult to understand why certain emotions are predicted based on visual or textual inputs, particularly in multimodal contexts, due to the interaction of multiple sources of information. High recognition accuracy alone is consequently not sufficient, as interpreting model behavior is essential for validating reliability and ensuring meaningful emotion analysis (Doshi-Velez and Kim, 2017). For example, vision-based models may depend on irrelevant background patterns instead of meaningful object features, leading to unreliable predictions (Lapuschkin et al., 2019). The present systems of multimodal emotion detection face challenges, primarily in achieving accuracy and explainability. Although the results of both unimodal and multimodal approaches on text, speech, and facial expressions have shown significant results individually, highlighting that it remains incapable of creating effective data fusion across these three modalities. The predictive reason behind specific emotions becomes difficult to comprehend because these decision-making systems lack transparency. Further, this fusion challenge is due to both modalities possessing different feature representations, temporal dynamics, and noise characteristics, which makes it difficult to extract complementary information (Geirhos et al., 2020).

To address these challenges, several studies use Explainable Artificial Intelligence (XAI) methods aiming to resolve the interpretability gap between complex AI models and human understanding. Explaining individual emotion

detection model predictions involves using LIME (Local Interpretable Model-Agnostic Explanations), SHAP (Shapley Additive exPlanations), and Grad-CAM (Gradient-weighted Class Activation Mapping). These XAI methods offer human-understandable explanations for model predictions. With these explanations, users gain important insights into the features influencing the model's decisions, increasing their trust in the models and helping to improve them. Some research studies have effectively combined these techniques to address the accuracy and interpretability challenges in emotion detection models Norval and Wang (2025), Lorch et al. (2025).

The study applies multimodal fusion, which further strengthens the interpretation outputs from text and image modalities to create a unified framework. By using weighted averaging along with concatenation methods or attention-based fusion to develop a powerful and complete emotion detection platform. By integrating multiple modalities, the system overall overcomes the limitations of unimodal approaches and effectively enhances emotion detection performance. To enhance this work further, we applied XAI methods post-hoc to offer qualitative insight into model decisions in both modalities. Image-based models (Grad-CAM) were used to localize emotion-related face regions Selvaraju et al. (2017), and text-based models (LIME) Ribeiro et al. (2016) were employed to identify the most significant tokens contributing to emotion prediction. This integration helps models rely on interpretable, meaningful patterns by promoting accuracy and consistency in model decisions aligned with human reasoning, thereby minimizing bias and increasing reliability in practical applications.

To systematically evaluate the effectiveness of multimodal emotion recognition, four experimental settings were designed. First, multimodal models were evaluated on text and image data to assess the accuracy of multimodal fusion for emotion classification. Second, unimodal experiments were conducted separately on text-only and image-only data to compare performance across modalities and analyze the individual contribution of each modality. Third, a systematic comparison was also performed between unimodal and multimodal approaches to determine which approach achieves high accuracy, demonstrating the effectiveness of multimodal integration over unimodal and single-modality approaches. Further, the XAI techniques were applied to the best-performing models from each setting, Grad-CAM for image-based models and LIME for text-based models, to generate interpretable explanations for emotion predictions. Fourth, the XAI outputs were manually reviewed and validated by a professional psychologist to confirm whether the highlighted visual and linguistic cues were grounded in meaningful emotional patterns rather than spurious artifacts. The main contributions of our work are as follows.

- **Unified Multimodal Framework** We propose ITMER, a multimodal emotion recognition framework that integrates image and text modalities to examine complementary visual and linguistic emotional cues for improved emotion classification.
- **Systematic Unimodal–Multimodal Comparison:** The framework conducts explicit experiments across text-only, image-only, and multi-

modal configurations, providing a direct analysis of each modality's individual contribution and the benefit of multimodal fusion, an evaluation rarely performed in prior work.

- **XAI-based Interpretability:** The framework incorporates explainable AI techniques, including Grad-CAM for image models and attention-based analysis for text models, to provide insight into model predictions and improve the interpretability of emotion recognition results.
- **Improved Performance over Unimodal Models:** Experimental results show that the proposed multimodal approach achieves high accuracy performance over unimodal baselines, achieving up to 92% accuracy and 92% macro F1-score, compared to 83.81% for image-only and 71% for text-only models, confirming the benefit of integrating multimodal fusion.

The rest of the paper is organized as follows: Section 2 presents the related work on emotion recognition, focusing on unimodal and multimodal approaches. Similarly, Section 3 presents our proposed pipeline for emotion recognition from text and images utilizing encoder-only, decoder-only, and encoder-decoder-only models. Moreover, Section 4 reports experimental and evaluation results for emotion recognition fine-tuning models on only text, only images, and combining both modalities. Similarly, explainability is calculated to identify the reason behind the emotion recognition made for each model. In Section 5, we have a comprehensive discussion about the proposed study, the limitations, and future work, and finally, Section 6 concludes the study.

2 Related Work

This section reviews prior research on emotion recognition utilizing deep learning algorithms and techniques across different benchmarks, i.e., textual, audio, image, and motion datasets. Emotion recognition (ER) is a critical research area within affective computing that has seen significant advancements in recent years, particularly in multimodal learning, deep learning, and human–computer interaction. The primary goal of ER is to accurately identify individuals emotion state through different channels, including textual conversation, facial expressions in videos, images, and speech signals. Due to the complexity and diversity of emotional expressions, emotion recognition tasks are highly dependent on effective multimodal data fusion that integrates complementary information from multiple data sources, i.e., text, speech, and visual cues. Recently, numerous studies have examined various fusion strategies and deep learning approaches for emotion recognition.

Visual emotion recognition focuses on identifying emotional states from images and videos and has advanced significantly with the development of computer vision and deep learning techniques. Emotions expressed in images are inherently abstract and subjective, making image-based emotion recognition

more challenging than text-based analysis. Deep learning methods, particularly CNNs and transformer architectures, have been widely adopted to extract discriminative features for emotion recognition in images Pereira et al. (2024). The shift to deep learning models, particularly Convolutional Neural Networks (CNNs), has demonstrated strong capability in learning complex facial patterns from large-scale datasets, often through robust feature fusion approaches like global-local joint representation learning Dong et al. (2022). Another study has demonstrated that CNNs outperform traditional SVM models on benchmark facial expression datasets, including CK+ and FER2013 (Ezerceli and Eskil, 2022). Lucey et al. (2010) and FER2013 Goodfellow et al. (2013). Nevertheless, CNN models tend to overfit when they are trained on relatively small datasets of facial expressions. To overcome these challenges, Knyazev et al. (2018) fine-tuned CNNs that were pretrained on a large face recognition dataset, and they obtained approximately 1% improvement over the previous result. In addition, Pons and Masip (2018) reported that FER could be performed at the same time with other related skills, such as landmark localization, and by that, this approach improves model results, indicating the benefits of multi-task learning in FER. Regarding Textual emotion recognition (TER), it has also seen significant progress, with recent research highlighting the effectiveness of transfer learning. For instance, Ameer et al. Ameer et al. (2023) fine-tuned pre-trained transformer models, including RoBERTa, XLNet, and DistilBERT, on multi-label emotion datasets in English and Chinese. Their best configuration, RoBERTa enhanced with a multiple-attention mechanism (RoBERTa-MA), achieved a state-of-the-art accuracy of 62.4%. This high-performing unimodal analysis establishes a benchmark for the text component of multimodal systems. Additionally, a study from SemEval focused on the use of deep learning approaches. The authors developed a multi-layer self-attention mechanism combined with a Bidirectional Long Short-Term Memory (Bi-LSTM) model. While this approach proved effective, BiLSTM models can struggle with long-range dependencies in text, whereas transformer-based models such as DeBERTa address this limitation through advanced attention mechanisms (Vaswani et al., 2017).

More recently, the BERT Model has achieved astounding performance in various NLP tasks. It is a deep multi-layer Transformer-based architecture that is trained in an unsupervised way on a large scale of unlabeled textual data to learn rich linguistic representations Devlin et al. (2019). Subsequently, it was followed by advanced pre-trained models like Transformer-XL and XLNet, which have also been introduced to enhance contextual understanding. Besides that, Cai et al. (2020) built a BERT model and integrated it with a Bi-LSTM to analyze and predict sentiment in user-generated text. A recent survey also discusses further examples of BERT-based methods to recognize text emotion Acheampong et al. (2021). Furthermore, a study used more than 144,000 tweets to create an emotion recognition model based on CNN, in which the Happy emotion was identified with a high rate of 73%, followed by Label anger with the lowest ratio of 37%. Park et al. (2020)

To overcome the limitations and ambiguities of unimodal approaches, multi-

modal fusion has gained attention in recent decades, as multimodal approaches have reported a significantly higher effectiveness as compared to unimodal approaches. As an example, Lan et al. (2020) revealed higher performance of the multimodal methods compared to the unimodal methods. Multimodal fusion approaches have been shown to improve emotion recognition performance by integrating diverse signals, including physiological and behavioral data, enabling richer and more robust affective representations Katada et al. (2022). Chandraumakantham et al. (2024), who proposed a novel Large Language Model (LLM)-based feature fusion strategy to overcome limitations in predicting emotions and handling emotional interplay across channels. Furthermore, the highest-performing contemporary models for text-image fusion often utilize modern transformer architectures: advanced systems combining BERT for text processing with Vision Transformers (ViT) for image analysis have been shown to achieve high accuracy, with Deshpande et al. Deshpande et al. (2025) reporting a classification accuracy of 96.91% in multimodal sentiment detection. More broadly, the entire landscape of multimodal emotion recognition has been systematically reviewed by Hosseini et al. (2024a), applying different approaches on audio, image, and text modalities. Their analysis highlighted that the successful integration of data relies on deep learning models and various fusion strategies (feature-level, decision-level, and hybrid) to enhance recognition rates. It also concluded that persistent challenges remain, including the subjective nature of emotion, cultural variability, and the reliance on extensive, annotated datasets.

The scarcity of large, labeled corpora often hinders the study of multimodal emotion recognition (MER). To overcome these challenges, Zhao et al. (2022)proposed MEmoBERT, a multimodal pre-training model designed to combine textual, visual, and acoustic representations from a large, self-collected corpus of unlabeled video data. The key focus of MEmoBERT was a prompt-based learning approach, which efficiently adapts the pre-trained Transformer to the classification task by reframing it as a masked text prediction problem. This strategy proved highly effective, achieving a new state-of-the-art Weighted Accuracy of 81.09% on the IEMOCAP benchmark.

Similarly, in the context of image–text sentiment analysis, recent works demonstrate the effectiveness of joint multimodal learning. For instance, Gong et al. Gong et al. (2024) proposed a multi-channel multimodal joint learning framework that integrates visual and textual features through cross-modal interaction mechanisms. Their approach was evaluated on the MVSA-Single and MVSA-Multiple datasets, achieving classification accuracies of 76.98% and 75.32%, respectively. Earlier multimodal emotion recognition approaches relied on conventional deep learning and simple fusion strategies. Recent work by Hazmoune Hazmoune and Bougamouza (2024) reviewed transformer-based multimodal emotion recognition models, including BERT, ViT, and cross-attention–based multimodal transformers, and showed that attention-driven architectures better capture cross-modal interactions and improve emotion recognition performance across multiple modalities.

2.1 Multimodal emotion fusion

Multimodal emotion recognition aims to integrate information from multiple signals, such as sound, language, and image in videos, for recognizing human emotions. Recent studies conducted employed multimodal fusion for different modalities. For instance, one study by Panchal et al. (2023)demonstrated that the method predicts user emotions by analyzing facial expressions in video input while simultaneously incorporating speech and text information. Experimental results show that emotion recognition methods combining multiple modalities significantly outperform single-modality approaches in terms of accuracy. Additionally, Chen et al. (2024) proposes a multimodal emotion recognition approach based on Graph Neural Networks (GNNs). This method represents the features of each modality as a node in a graph and uses the edges in the graph to capture the relationships between modalities. This approach effectively handles the complex interactions between different modalities and demonstrates promising results in emotion recognition in conversations. In contrast to multimodal fusion, Hosseini et al. (2024b) integrates audio, video, and textual data by incorporating a multimodal emotion ecognition method. The authors do not perform direct trimodal fusion, but first combined audio-video and audio–text features independently, and subsequently fused the outputs using multinomial logistic regression. A chi-square test is done during feature selection in an attempt to reduce redundancy as well as suppress noise before the classification process. Comparison of the IEMOCAP dataset demonstrated that the proposed multimodal fusion method achieved an accuracy of 82.9%, outperforming several unimodal and existing multimodal approaches.

Further, other studies have emphasized adaptive fusion strategies for multimodal emotion recognition. Lu et al.Lu et al. (2025) proposed a transformer-based multimodal fusion framework that integrates text, speech, and visual modalities for emotion recognition. The model achieved an accuracy of 74.98% on IEMOCAP (weighted F1: 75.11%) and 67.93% accuracy on MELD, outperforming both unimodal and bimodal baselines. However, the framework incurs a high computational cost and depends on large-scale multimodal data, limiting its applicability in low-resource scenarios.

Additionally, a recent study on multimodal fusion utilized Transformer-based models for textual and audio data. Jatoth et al. (2025)explored multimodal emotion detection using a suite of Transformer-based models, including BERT and RoBERTa, alongside audio models like AST and Wav2Vec2. Crucially, their findings demonstrated that multimodal fusion consistently outperforms unimodal models. The findings further show that Late Fusion (Text + Audio) is particularly effective, achieving the highest weighted F1-Score of approximately 78% on IEMOCAP. For comparison, their best-performing unimodal text models (RoBERTa+BiLSTM and BERT+BiLSTM) achieved 73.33% and 72.75%, respectively.

However, the above studies demonstrate the effectiveness of multimodal fusion for emotion recognition tasks. Most existing approaches rely primarily on the IEMOCAP dataset, focusing on video-based facial expressions and con-

versational speech or text, thereby limiting their applicability to social media–based textual and image data. In particular, Jatoth et al. (2025). and Zhang et al. (2023) focused only on bimodal fusion settings, lacks a systematic unimodal–multimodal performance comparison, and does not evaluate modality effectiveness independently. Moreover, their studies focus on sentiment or constrained emotional contexts, thereby offering limited analysis of emotion recognition in unconstrained social media environments.

In contrast, this study employs multimodal fusion over social media text and images across six emotion classes. Furthermore, separate unimodal experiments are conducted on both textual and visual modalities to explicitly analyze modality contributions and fusion effectiveness, addressing the gaps identified in prior work.

3 Methodology

3.1 Problem Definition

This study addresses a six-class emotion classification problem with the following emotion categories: Happy, Surprise, Angry, Sad, Neutral, and Fear. Each input sample is assigned a single emotion label encoded numerically as: Happy (0), Surprise (1), Angry (2), Sad (3), Neutral (4), and Fear (5). The objective is to automatically predict the correct emotion class using textual information, visual information, or a combination of both modalities. To systematically analyze the contribution of each modality, four experimental settings were designed: text-only emotion classification, image-only emotion classification, Multimodal emotion classification, and Explainable AI (XAI) integration.

The proposed methodology consists of three main modules: i) Text-Only emotion classification, ii) Image-Only emotion classification, and iii) Multimodal emotion classification, as shown in Figure 1. In the Text-only classification module, we have employed encoder-only models like BERT(Bidirectional Encoder Representations from Transformers), DistilBERT, and RoBERTa (Robustly Optimized BERT Pretraining Approach). Similarly, in the second module for image-based emotion classification, ResNet, EfficientNet, and DenseNet CNN-based architecture models were employed. In the third module, we utilized a multimodal approach to identify emotion from text as well as image, where we combined an encoder-only model with CNN-based models. Each of the modules is comprehensively presented in the following section.

3.2 Dataset Collection

To support unimodal and multimodal emotion recognition, data were collected from multiple publicly available emotion datasets containing textual data, visual data, and aligned text-image pairs. Since no single dataset fully satisfied in all experimental settings, a unified dataset was constructed by integrating samples from different sources. Three emotion classes, *Happy*, *Surprise*, and *Sad*, were

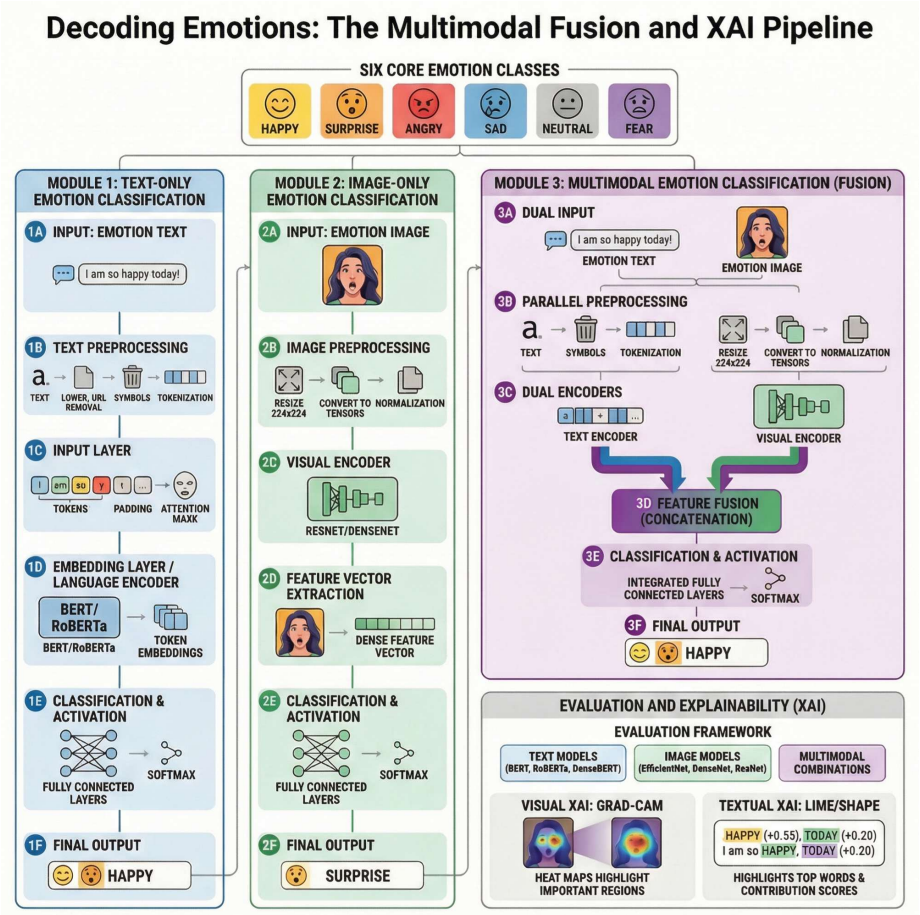


Figure 1: Overview of the proposed high-level workflow for multimodal emotion recognition

collected from the same source. The remaining three emotion classes were collected from different publicly available datasets hosted on Kaggle. Textual data for the *Happy*, *Surprise*, and *Sad* classes were obtained from publicly available Kaggle datasets. *Angry*, *Neutral*, and *Fear* textual datasets were also collected from different Kaggle sources.

All emotion labels from the original datasets were mapped to a unified six-class emotion taxonomy (Happy, Surprise, Angry, Sad, Neutral, and Fear). Samples with incompatible or ambiguous labels were excluded. For multimodal experiments, text–image pairs were either directly taken from existing multimodal datasets or constructed by pairing text and image samples that shared the same emotion label. 2 presents the label distribution of the unified emotion dataset across six emotion categories. The rationale for using this pairing

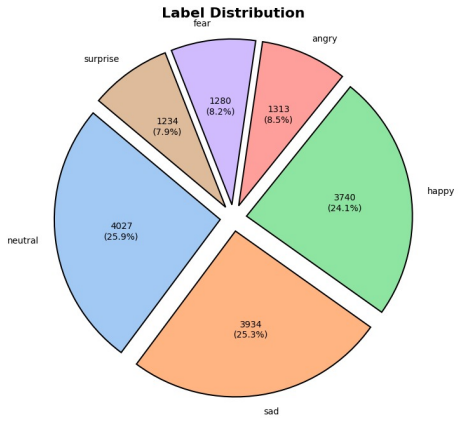


Figure 2: Label distribution of six emotion classes in the unified dataset (Total N = 15,528)

strategy was to ensure balanced class representation and similar emotion labels across modalities, enabling controlled assessment. Although not all pairs were initially co-occurring in real-world posts, this design enables a systematic analysis of modality contributions within the ITMER framework. The image datasets were collected from multiple publicly available sources on Kaggle, as well as through manual curation, because it was not feasible to obtain the required number of samples for all emotion classes from a single source. The corresponding dataset links are provided on Kaggle.
[1] [2] [3]

3.3 Module 1: Text Only Emotion Classification

3.3.1 Model Input

The input to the text-based emotion classification module consists of short textual statements from social media post datasets. Each text instance is a user-generated message associated with one of the 6 predefined emotion classes: Happy, Surprise, Angry, Sad, Neutral, and Fear. The textual samples vary in length and are written in informal language, using emojis, punctuation, and web links; they require preprocessing before being passed to the language encoder models.

[1] Textual datasets available at: `https://www.kaggle.com/datasets/nelgiriyewithana/emotions`; `https://www.kaggle.com/datasets/subrahmanya090/expression`; `https://www.kaggle.com/datasets/simaanjali/emotion-analysis-based-on-text`.

[2] Image datasets available at: `https://www.kaggle.com/datasets/yousefmohamed20/sentiment-images-classifier`; `https://www.kaggle.com/datasets/isrcoimbra/emotionnet-6-realistic-facial-emotion-dataset`; `https://www.kaggle.com/datasets/aryan0910/face-emotion-dataset-emotionx-09`.

[3] Mendeley dataset available at: `https://data.mendeley.com/datasets/zfcd4bny82/1`.

3.3.2 Text Preprocessing

Textual data were preprocessed to reduce noise and improve model performance. The preprocessing steps included converting to lowercase, removing URLs, punctuation, and emojis, and removing non-alphanumeric characters, followed by tokenization. All text sequences were padded or truncated to a fixed length to ensure uniform input dimensions during model training.

3.3.3 Language Encoder Models

Transformer-based language encoders were pretrained and applied to obtain contextual representations of textual inputs. Precisely, the encoder-only models were BERT, RoBERTa, and DistilBERT, which were created to generate dense sentence embeddings in the form of a sequence of tokenized text. The aggregated textual feature represented by the [CLS] token representation was moved to a fully connected classification layer with the softmax activation function to estimate the emotion class.

3.4 Module 2: Image Only Emotion Classification

3.4.1 Model Input (Image Module)

Facial emotion images were set as input for the image-only emotion classification module. Each image contained an expression of one of the six emotion categories. The images were processed separately and run through a visual encoder to extract discriminative features based on facial expressions and contextual visual features.

3.4.2 Image Preprocessing

All images were preprocessed using a standardized transformation pipeline. Specifically, each image was resized to a spatial resolution of 224×224 pixels and then converted into a tensor representation using pixel intensity values scaled to the range [0, 1]. The image preprocessing pipeline is defined as described above.

3.4.3 Visual Encoder Models

In order to extract visual features of emotional expressions, several deep convolutional neural networks were used, such as ResNet, DenseNet-121, and EfficientNet-B0. Such architectures can represent hierarchical spatial codes of low-level facial representations up to high-level emotional features. The output feature encoder of the visual encoder is further processed to fully connected layers to make emotion classification.

3.5 Module 3: Model Fusion-Based Emotion Classification Model

3.5.1 Dual Input (Multimodal)

The multimodal module was set with textual and visual inputs corresponding to the same instance of emotion. Transformer-based language encoders processed the textual data, and the associated images are processed through convolutional neural networks. This structure with two inputs enables the model to have the complementary emotion cues of both modalities.

3.5.2 Parallel Preprocessing

Preprocessing of text and image inputs is carried out in parallel and followed by feature extraction. Textual inputs are tokenized, padded, and encoded with the corresponding language model tokenizer. Simultaneously, image inputs are resized and normalized through the image preprocessing pipeline. This is parallel preprocessing, ensuring that both modalities are ready for feature extraction by their respective encoders.

3.5.3 Dual Encoders

Two independent encoders process each modality in the multimodal architecture. A language encoder based on a transformer that extracts contextual textual embeddings, and a convolutional neural network that uses image feature encodings. These modality-specific representations capture complementary semantic and visual information useful for emotion recognition.

3.6 Models Evaluation and Integration of Models Explainability (XAI)

3.6.1 Language Models Explainability

Local Interpretable Model-agnostic Explanations (LIME) were employed to interpret the predictions of text-based models. LIME assesses the measurements of effects of individual tokens by generating perturbed text samples and tracking changes in model predictions. The result descriptions highlight the significant words that contribute to the expected emotional class.

3.6.2 Visual Models Explainability

To improve the interpretability of image-based emotion recognition, two post-hoc explainability techniques were employed: Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME). Both methods were applied to trained vision models to analyze which image regions contributed to the predicted emotion classes.

Grad-CAM: Grad-CAM was used to generate class-discriminative localization maps by leveraging gradients from the final convolutional layer. For a

given input image x and predicted class c, the gradient of the class score y^c with respect to the activation maps A^k, the target convolutional layer was computed. The channel-wise importance weights α_k^c were obtained through global average pooling of the gradients:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k},$$

where Z represents the total number of spatial locations in the feature map. The Grad-CAM heatmap was then computed as a weighted combination of the activation maps, followed by a ReLU operation:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right).$$

The heatmap was resized to the input resolution and normalized to the range $[0, 1]$ to visualize it. This enables qualitative analysis of the facial region, which contributed most to the significant emotion predictions.

LIME: In addition to Grad-CAM, LIME was applied to provide local, model-agnostic explanations. LIME generates interpretable explanations by perturbing the input image into multiple samples and observing the corresponding model prediction changes. A sparse linear surrogate model is fitted around the specific input instance, and the most influential superpixel regions contributing positively to the predicted class are highlighted. In this study, LIME explanations were generated using 1000 perturbed samples per input image, and the top contributing regions were visualized to support the interpretability analysis of the image-based classifiers.

4 Experiments, Results, and Evaluation Strategy

4.1 Evaluation Metrics

To evaluate the performance of the classification model, we used standard metrics derived from the confusion matrix. The confusion matrix consists of four outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Here, TP represents positive samples correctly classified as positive, TN represents negative samples correctly classified as negative, FP represents negative samples incorrectly classified as positive, and FN represents positive samples incorrectly classified as negative.

4.1.1 Accuracy

Accuracy measures the overall proportion of correct predictions among all predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Accuracy provides a general measure of model correctness; however, it may be misleading in imbalanced datasets where one class dominates.

4.1.2 Precision

Precision measures the proportion of predicted positive instances that are truly positive:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Precision is particularly important when the cost of false positives is high.

4.1.3 Recall (Sensitivity)

Recall, also referred to as sensitivity, measures the proportion of actual positive instances that are correctly identified:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Recall becomes critical in applications where false negatives are costly, such as missing abnormal findings in radiology reports.

4.1.4 F1-score

F1-score is the harmonic mean of precision and recall, to provide a balanced measure of results when false positives and false negatives are both important:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

This measure is specifically applicable to class imbalance and where a trade-off is required between precision and recall. The comparison of the model performance focuses on the macro-averaged measures because emotion recognition is multi-class and can be unbalanced. The Macro-F1 offers a balanced-weighted evaluation of all types of emotion categories, and the majority of classes predominate the overall performance measurement.

4.1.5 Text-Only Emotion Classification

For text-only emotion classification, multiple transformer-based pretrained language models were evaluated, including DistilBERT, RoBERTa, BERT, and T5. These models were used to encode textual inputs into contextualized semantic representations. The extracted embeddings were then passed through task-specific fully connected layers, followed by a softmax output layer to predict probabilities across the six emotion classes. Dropout regularization was applied during training to reduce overfitting and improve generalization. Performance comparisons among the evaluated models are reported in the Results section.

4.1.6 Image-Only Emotion Classification

The image-based emotion classification model employs a deep convolutional neural network (CNN) to extract the discriminating visual features linked to emotion expression and other contextual cues. Pre-trained weights were used to initialize the CNN to leverage transfer learning. The features extracted were fed to the fully connected layers, and a softmax classifier was used to predict the emotion label.

4.1.7 Multimodal Emotion Classification

The multimodal model combines both textual and Image data information. Feature Embeddings of the text encoder and the image encoder were concatenated to create a unified multimodal representation. A combination of these feature vectors was subsequently subjected to further dense layers that were intended to learn cross-modal interactions. The last softmax layer outputs the predicted emotion category among the six possible categories.

4.2 Training Strategy

Training was done by a uniform optimization strategy on all models to make them comparable to each other across modalities. The Adam optimizer was adopted, and categorical cross-entropy loss was adopted as the optimization method. Validation loss was used to early stop to avoid overfitting. Empirically chosen hyperparameters included the batch size and the number of epochs, which remained constant across all the experiments.

4.3 Explainable AI (XAI) Integration for Text-based Models

To interpret the predictions of the text-only emotion recognition model, explainable AI was applied using the Local Interpretable Model-agnostic Explanations (LIME) technique. The text classifier was based on a transformer encoder (DistilBERT) fine-tuned for six-class emotion classification. After training, the predictor returned both predicted emotion labels and probability scores for each class Ribeiro et al. (2016).

LIME provides local explanations by approximating the behavior of a complex model around a specific input instance. Given an input text sample, multiple perturbed versions of the text are generated by removing or masking selected tokens. The trained model is then used to obtain prediction probabilities for these perturbed samples. Based on this local neighborhood, LIME fits a sparse linear surrogate model to estimate the contribution of individual words toward the predicted emotion class. The output explanation highlights the most influential words along with their positive or negative contribution weights.

In this study, LIME explanations were generated for selected test samples by extracting the top contributing tokens. ($n = 10$) for the predicted emotion

class. These token-level importance scores were used to qualitatively analyze which words and contextual phrases influenced the model's predictions, thereby improving the interpretability of the transformer-based emotion recognition outputs.

4.4 Experimental Setup

All experiments were conducted on an NVIDIA A100 GPU (80 GB GPU RAM, 167 GB system RAM). The text-only RoBERTa model was implemented using the ktrain library with TensorFlow backend, while the image-only DenseNet-121 and multimodal RoBERTa + ResNet18 models were implemented in PyTorch. Across all three settings, a batch size of 16 and a learning rate of 2e-5 were used. The text-only and multimodal models were trained using the AdamW optimizer, while the image-only DenseNet-121 model was trained using the Adam optimizer. CrossEntropyLoss was used as the loss function across all experiments. Text inputs were tokenized with a maximum sequence length of 400 tokens for the text-only model and 256 tokens for the multimodal model. All images were resized to 224×224 pixels and normalized before training. The text-only model was trained for 3 epochs using a 70/20/10 train/validation/test split, the image-only model was trained for 5 epochs with an 80/20 train/test split, and the multimodal model was trained for 3 epochs with an 80/20 train/test split. All preprocessing pipelines and evaluation protocols were kept consistent across modalities to ensure fair and reproducible comparison. 1

Table 1: Experimental setup and hyperparameters for best-performing models across modalities.

Parameter	**RoBERTa (Text-only)**	**DenseNet-121 (Image-only)**	**RoBERTa + ResNet18 (Multimodal)**
Framework	TensorFlow / ktrain	PyTorch	PyTorch
Optimizer	AdamW	Adam	AdamW
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch Size	16	16	16
Epochs	3	5	3
Loss Function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Max Sequence Length	400 tokens	–	256 tokens
Image Size	–	224×224	224×224
Train/Val/Test Split	70/20/10	80/–/20	80/–/20
Hardware	NVIDIA A100 GPU (80 GB GPU RAM, 167 GB System RAM)		

5 Results

5.1 Performance of Multimodal Fusion Models

Section 4 presents a systematic evaluation of the proposed ITMER framework across text-only, image-only, and image–text multimodal emotion recognition settings. ITMER is designed as a unified evaluation framework rather than a

single novel architecture, enabling fair comparison of unimodal and multimodal models under consistent preprocessing and evaluation protocols.

The performance of the evaluated multimodal fusion architectures for emotion recognition is summarized in Table 2. All models were assessed on an independent test set (N = 300) using accuracy, macro-averaged, and weighted-averaged precision, recall, and F1-score as well as weighted evaluation metrics to account for potential class imbalance.

The proposed RoBERTa + DenseNet-121 model achieved an accuracy of 90.00%, with a macro F1-score of 92.00% and a weighted F1-score of 90.00%. This configuration also attained the highest macro recall (92.00%) among the transformer–CNN fusion models, indicating balanced performance across emotion categories.

The BERT + ResNet architecture achieved the same overall accuracy (90.00%) as the proposed approach, with comparable macro-precision (91.00%) and macro-recall (92.00%). However, its macro F1-score (91.00%) was marginally lower, suggesting a slightly reduced balance across classes.

The DistilBERT + EfficientNet-B0 configuration obtained an accuracy of 88.00% and a macro F1-score of 88.00%. Although competitive, DistilBERT's lower model capacity relative to full transformer architectures appears to slightly limit overall multimodal fusion performance.

The EfficientNet-B0 + GPT-2 model achieved an accuracy of 90.93%, with macro and weighted F1-scores of 90.73% and 88.65%, respectively. While macro performance remains competitive, the lower F1-score weight indicates reduced robustness to class imbalance.

Among baseline multimodal models, FLAVA demonstrated substantially lower performance, achieving 69.29% accuracy and a macro F1-score of 66.28%. In contrast, CLIP-Text + ViT-B/32 performed competitively, achieving 89.00% accuracy and a macro F1-score of 91.00%, confirming the effectiveness of large-scale vision–language pretraining for multimodal emotion recognition tasks.

The SigLIP model achieved the highest overall accuracy (92.00%); however, its macro recall (86.00%) was lower than that of the proposed approach, indicating less consistent performance across emotion classes despite strong overall accuracy.

Additional multimodal configurations were considered to evaluate the performance of transformer CNN fusion approaches. The RoBERTa + ResNet18 model achieved an accuracy of 91.00%, with macro precision and recall of 92.00% and 93.00%, respectively, and a macro F1-score of 92.00%. It was shown that this setup achieved high, well-balanced performance across all emotion classes, indicating the effectiveness of integrating contextual textual representations from RoBERTa with deep residual visual feature extraction.

BERT + DenseNet-121 models were shown to have competitive results, achieving an overall accuracy of 90.00% and a macro F1-score of 91.00%. The visual encoder has dense connectivity, enabling improved reuse of features and gradient flow, which appears to contribute to the recognition of emotions performance across several different categories.

The RoBERTa + EfficientNet-B0 model had an accuracy of 87.00% with

a macro F1-score of 89.00%. Despite being slightly lower than the highest-performing models, the architecture still had quite high recall values, suggesting that EfficientNet-based visual encoders can effectively capture relevant face and contextual emotional cues in combination with transformer-based text representations.

Similarly, the BERT + EfficientNet-B0 model had an overall accuracy of 87.00% and equalized macro precision and recall values of 89.00%. These results suggest that even though the EfficientNet-based multimodal combinations are competitive, architectures incorporating deeper residual networks or densely connected networks are likely to offer a slightly higher and more stable multimodal representation. In general, the findings indicate that transformer-CNN multimodal fusion architectures consistently outperform unimodal approaches. Models combining strong textual encoders (e.g., BERT or RoBERTa) and deep convolutional visual feature extractors (e.g., DenseNet or ResNet) provide the most balanced and robust performance across emotion categories.

Emotion prediction patterns further emphasize the advantages of multi. modal fusion. Emotions like *happy* and *fear* are usually recognized with higher accuracy because of their clear visual and textual cues. In contrast, emotions such as *neutral* and *sad* show more confusion, probably due to overlapping semantic and contextual features. These findings match common difficulties in multimodal emotion recognition, where subtle affective states often have similar linguistic and visual traits.

Table 2: Performance comparison of multimodal fusion architectures.

Classifier	**Accuracy**	**Precision**		**Recall**		**F1-Score**	
	(%)	Macro	Weighted	Macro	Weighted	Macro	Weighted
RoBERTa + DenseNet-121 (Ours)	90.00	91.00	90.00	92.00	90.00	92.00	90.00
DistilBERT + EfficientNet-B0	88.00	89.00	88.00	90.00	88.00	89.00	87.00
BERT + ResNet-50	90.00	91.00	90.00	92.00	90.00	91.00	90.00
EfficientNet-B0 + GPT-2	88.83	90.93	89.20	91.07	88.83	90.73	88.65
FLAVA (Baseline)	69.29	63.08	59.60	75.74	69.29	66.28	60.29
CLIP-Text + ViT-B/32 (Baseline)	89.00	91.00	90.00	91.00	89.00	91.00	89.00
SigLIP	92.00	92.00	89.00	86.00	88.00	88.00	88.00
RoBERTa + ResNet18	91.00	92.00	91.00	93.00	91.00	92.00	91.00
BERT + DenseNet-121	90.00	91.00	90.00	92.00	90.00	91.00	90.00
RoBERTa + EfficientNet-B0	87.00	88.00	87.00	91.00	87.00	89.00	87.00
BERT + EfficientNet-B0	87.00	89.00	87.00	89.00	87.00	89.00	87.00

5.2 Performance of Uni-modal Text-Based Models

The performance of uni-modal (text-only) emotion classification models evaluated on the test set (N = 3,102) is summarized in Table 3. Overall, all text-based approaches achieved notably lower performance compared to the multimodal fusion architectures presented earlier.

The performance of unimodal (text-only) emotion classification models evaluated on the test set (N = 3,102) is summarized in Table 3. Overall, the

performance of all text-based methods was particularly lower than that of the multimodal fusion models described above, emphasized on the limitations of using solely textual cues for emotion recognition.

Among transformer-based models, BERT, RoBERTa, and DistilBERT achieved almost the same accuracy values of 71.00%, with macro F1-scores ranging between 69.00% and 70.00%. These models exhibited relatively high macro recall (78.00%), indicating that they successfully identify a large proportion of emotion instances. However, the comparatively lower macro precision suggests an increased rate of false positives across emotion categories.

The T5 model demonstrated the weakest performance among the evaluated unimodal architectures, achieving an accuracy of 66.00% and a macro F1-score of 65.00%. This result indicates that generative text-to-text architectures may be less effective for fine-grained emotion classification when trained in a purely text-based setting.

The GPT-2-based configuration achieved an accuracy of 70.87%, with a macro F1-score of 69.70% and a weighted F1-score of 64.58%. The disparity between macro and weighted metrics suggests sensitivity to class distribution, where performance varies across different emotion categories.

Table 3: Performance comparison of uni-modal (text-only) emotion classification models on the test set (N=3102).

Model	**Accuracy**	**Precision**		**Recall**		**F1-Score**	
	(%)	Macro	Weighted	Macro	Weighted	Macro	Weighted
DistilBERT	71.00	69.00	67.00	78.00	71.00	69.00	63.00
RoBERTa	71.00	69.00	67.00	78.00	71.00	69.00	63.00
BERT	71.00	68.00	67.00	78.00	71.00	70.00	64.00
T5	66.00	61.00	58.00	73.00	66.00	65.00	59.00
GPT-2 + EfficientNet-B0	70.87	67.03	65.39	77.90	70.87	69.70	64.58

5.2.1 Performance of Uni-modal (image-only) Emotion Classification

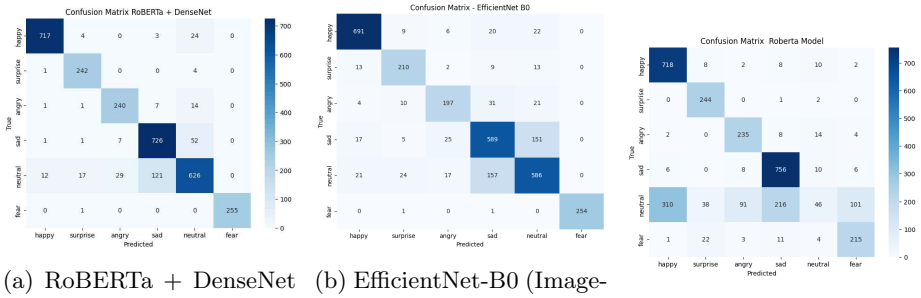
The performance of uni-modal (image-only) emotion classification models was evaluated on the test set (N = 3,106), summarized in Table 4. Overall, image-based approaches showed stronger performance and outperformed text-only models. However, they remain less effective than multimodal fusion architectures. DenseNet-121 achieved the highest performance Score, with an accuracy of 83.81% and a macro F1-score of 84.76%, among the evaluated visual models, indicating DenseNet-121 strong ability to capture discriminative visual features related to emotional expressions. ResNet and EfficientNet-B0 also exhibited competitive performance with an accuracy of 82.03% and 81.36%, respectively, with balanced precision and recall across emotion categories.

In addition, the Vision Transformer (ViT) model achieved comparatively lower performance, with an accuracy of 80.68% and a macro F1-score of 82.52%.

This suggests that convolutional neural network architectures remain more effective than transformer-based models for image-only emotion recognition in the evaluated setting. In general, these findings suggest that while unimodal image-based models achieve stronger performance than unimodal text-only approaches (71% accuracy), they remain below the multimodal fusion architectures, which achieve up to 92% accuracy. This performance gap underscores the significance of jointly modeling visual and linguistic cues, as multimodal integration enables a more comprehensive and robust representation of emotional expressions than either of the two modalities. 3 shows the confusion matrices of the three best-performing models.

Table 4: Performance comparison of uni-modal (image-only) emotion classification models on the test set (N=3106).

Model	**Accuracy**	**Precision**		**Recall**		**F1-Score**	
	(%)	Macro	Weighted	Macro	Weighted	Macro	Weighted
Vision Transformer (ViT)	80.68	85.59	82.12	81.12	80.68	82.52	80.37
EfficientNet-B0	81.36	83.39	81.40	83.19	81.36	83.27	81.36
DenseNet-121	83.81	84.89	83.74	84.72	83.81	84.76	83.73
ResNet	82.03	84.33	82.23	83.04	82.03	83.63	82.09
MobileNetV2	78.72	81.51	78.90	79.61	78.72	80.47	78.75



(a) RoBERTa + DenseNet (Multimodal) (b) EfficientNet-B0 (Image-only) (c) RoBERTa (Text-only)

Figure 3: Confusion matrix of the best-performing models across modalities: (a) multimodal model RoBERTa+DenseNet, (b) image-only model EfficientNet-B0, and (c) text-only model RoBERTa.

5.3 Explainable AI Results for Image-based Models

Grad-CAM heatmaps were applied to the test samples to analyze the decision-making behavior of the image-only emotion classification models. The outcomes of the generated heatmaps indicate that the model consistently focused on emotion-relevant facial regions when making predictions. In the majority of samples, higher activation is found in most instances around the prominent

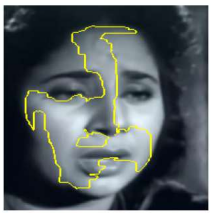
components of the face, such as eyes, mouth, and eyebrows, suggesting that the model is dependent on meaningful facial cues rather than irrelevant background regions. However, in a small number of challenging cases, activation was also observed in non-facial regions such as background areas and lighting artifacts, which may have contributed to misclassification in those instances.



(a) Original (True:3 Pred:3)



(b) Original (True:0 Pred:0)



(c) LIME



(d) LIME



(e) Grad-CAM



(f) Grad-CAM

Figure 4: Explainable AI visualizations for emotion recognition using LIME and Grad-CAM.

In addition, LIME was employed on selected image samples to identify local superpixel regions contributing positively to the predicted emotion class. To compute LIME explanations on every input image, 1000 perturbation samples were visualized to support the analysis of interpretability. The regions identified by LIME across various test samples were largely consistent with the facial regions highlighted by Grad-CAM, indicating consistency between the two complementary explanation methods. This correspondence between gradient-based and model-agnostic explanations provides additional qualitative evidence of image-based models focused on emotion-relevant facial features, thereby enhancing the general reliability and transparency of the proposed framework.

However, in certain challenging samples, the explanation maps show attention to non-facial regions such as background elements or lighting artifacts, which may have led to occasional misclassifications in those specific instances. Despite these limitations, the visual explanations outputs confirm that the image-based models effectively learned emotion-relevant facial patterns. Furthermore, these findings highlight potential directions for future research, including incorporating face-cropping preprocessing or attention-guided training to reduce model sensitivity to irrelevant background regions. The XAI visualizations for the image-based emotion recognition are shown in Fig. 4.

5.4 Explainable AI Results for Text-based Models

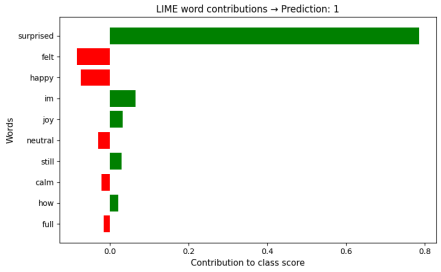
To interpret the predictions of the text-based emotion classification model, Local Interpretable Model-agnostic Explanations (LIME) was applied to the fine-tuned DistilBERT classifier. The explanations were generated using the model's probability outputs (predict-proba) through a custom prediction function, ensuring compatibility with LIME. For each selected test instance, LIME was configured to extract the top 10 most influential textual features (num-features = 10) contributing to the predicted emotion class. These features were assigned weights indicating their positive or negative influence on the model's decision. The LIME analysis reveals that the model assigns strong positive weights to emotionally explicit words. For instance, tokens such as "happy," "joy," and "excited" contribute positively toward the happy class, while words like "sad," "lost," and "angry" push predictions toward their respective negative emotion classes.

This demonstrates that the DistilBERT model effectively captures semantically meaningful emotional cues. A key observation from the experiments is the model's behavior on complex sentences containing multiple emotional states. For example, in the test sentence:

"I felt happy and full of joy when I got the news, but later I was sad thinking about what I lost, and now I'm just neutral and calm, though I'm still surprised," LIME highlights competing contributions from different tokens: Positive weights for "happy" and "joy." Negative or opposing contributions from "sad" and "lost." Moderate contributions from "neutral" and "surprised." This indicates that the model considers multiple emotional cues simultaneously, leading to a final prediction based on the dominant aggregated signal. Such behavior explains occasional ambiguity or reduced confidence in predictions for multi-emotion inputs. The explanations were visualized using horizontal bar plots, where Positive weights (green bars) indicate words supporting the predicted class, and negative weights (red bars) indicate words opposing the prediction.

Additionally, 5. LIME highlighted text visualization provides a complementary multi-emotion test sentence where the model's final prediction corresponds to class 1 (surprise). As we can see in the figure, words like "surprised" strongly dominate the prediction, with a significant positive score of approximately 0.79, displayed as a prominent green bar. However, competing tokens such as "felt," "happy," "neutral," and "calm" have negative weights (red bars), demonstrat-

ing that they oppose the predicted class. Although words like “im”, “still”, and ”how” exhibit minimal and positive contributions, whereas”full” has a slight negative weight. The visual comparison between the contribution of ”surprised” and all remaining words shows that the model’s prediction is influenced predominantly by a single high-confidence token, despite the sentence containing cues from several emotion categories.



(a) Explanation results of LIME-based DistilBERT on a multi-emotion input.



(b) LIME explanation of DistilBERT: contributions of words, highlighting tokens, and prediction probability.

Figure 5: LIME-based Results Explanation of the DistilBERT Text Emotion Classifier on a Multi-Emotion Test Instance.

5 Sub figure (a) shows the LIME explanation panel for the given test instance, which consists of three components. The left panel shows the prediction probability distribution among emotion classes, where class 1 achieves the highest probability of 0.76, followed by class 4 at 0.17, then class 0 at 0.05, class 5 at 0.01, and all remaining classes at 0.00. His distribution supports the fact that although the model predicts class 1 with high confidence, a non-negligible probability mass is assigned to class 4, indicating the sentence’s inherent emotional complexity. The center panel represents the word contribution bar chart in an interactive LIME format, which shows the contribution scores of the top features compared to class 1 versus NOT 1, with the highest positive score of 0.79. for class ”surprised”. The input sentence is displayed in the right panel with the most influential token, ”surprised,” highlighted in orange directly in the sentence. ”I felt happy and full of joy when I got the news, but later I was sad thinking about what I lost, and now I’m just neutral and calm, though I’m still surprised by how quickly everything changed today.”

This embedded highlighting supports that the model’s attention is connected to the word “surprised” and provides a human-readable description of the classification decision intuitively. Further, these visualizations provide intuitive insight into how individual words influence model decisions and allow easy identification of dominant emotional drivers in the text. Beyond isolated keywords, LIME results suggest that the model leverages contextual meaning captured by transformer embeddings. Words appearing in emotionally rich contexts receive higher importance scores compared to neutral or filler words, confirming that the model is not solely keyword-dependent. In certain cases, LIME reveals

that the model assigns importance to less relevant tokens or fails to correctly interpret nuanced expressions, such as implicit emotions or subtle transitions between emotional states.

This is particularly evident in sentences with sarcasm or gradual emotional shifts, where contributions are distributed across multiple classes. Overall, the LIME-based analysis confirms that the DistilBERT model makes decisions based on interpretable and semantically meaningful textual features, while also highlighting its limitations in handling multi-emotion and contextually complex inputs. This strengthens the transparency and reliability of the proposed multimodal framework.

6 Discussion

Human-computer interaction requires enhanced emotional awareness and transparency. In contribution to that, this study evaluated a multimodal emotion recognition framework integrated with Explainable AI (XAI) mechanisms. The results of the proposed work with XAI showed notable improvement, demonstrating that multimodal fusion significantly improves the performance of emotion classification over unimodal approaches. Based on these results, the multimodal architectures consistently outperformed unimodal models, achieving 92% accuracy and a macro-F1 of 92% for the RoBERTa + ResNet18 configuration, while the text-only model DistilBERT had a maximum accuracy of 71%, and the image-only model DenseNet-121 had an accuracy of 83.81%. These findings align with a larger body of literature confirming that multimodal emotion recognition systems offer a deeper and more robust analysis of emotional states compared to unimodal approaches. For instance, a recent study such as Deshpande et al. (2025) reported comparable high classification accuracy in multimodal image–text sentiment detection frameworks.

A key contribution of this work is the comparison between modalities. The results revealed that image-only models outperformed text-only models. Text models exhibited higher recall but lower precision, while image models provided more discriminative feature representations. This suggests that visual modality provides stronger discriminative features for emotions. The findings align with the existing work done on MER. Unimodal systems may misinterpret emotions as they rely solely on a single source of signals for emotional interpretation, which often results in decreased accuracy, especially in dynamic environments where multimodal data provides complementary cues for a better understanding. Nguyen et al. (2026) The present study directly addresses this limitation by demonstrating that fusing textual and visual representations mitigates modality-specific weaknesses and produces a more balanced precision–recall tradeoff across all six emotion categories. This explicit unimodal–multimodal comparison is rarely performed in prior work on image–text emotion recognition, and thus strengthens the analytical depth of the ITMER framework.

Among multimodal models, RoBERTa + DenseNet-121 achieved strong macro performance. SigLIP achieved the highest accuracy but lower macro recall.

FLAVA underperformed significantly. This suggests that: Simple feature concatenation with strong encoders can outperform complex unified multimodal pretraining models. Cross-modal alignment quality matters more than model complexity. Large pretrained contrastive models (CLIP/SigLIP) generalize well but may suffer from class imbalance sensitivity.

The integration of Grad-CAM and LIME significantly enhances the transparency of this framework. Image explanations highlight emotion-relevant facial areas, i.e., the eyes, mouth, and eyebrows. While the Text explanations revealed that emotion-related tokens dominate the prediction score. The convergence of both approaches on the same facial regions gives additional confidence in the reliability of the visual explanations. However, in complex inputs, for example, sentences with multiple emotions or images with complex backgrounds. The contributions score was divided across competing features, suggesting sensitivity to irrelevant or ambiguous inputs.

6.1 Limitations

Despite strong performance, several limitations exist:

Text-image pairs were partially constructed rather than naturally co-occurring. Dataset size remains moderate. The fusion method relies primarily on concatenation. The computational cost of multimodal transformers is high. Future work may explore: Cross-attention fusion mechanisms, Real-world aligned multimodal datasets, Lightweight multimodal architectures, Domain adaptation across social media platforms.

7 Conclusion

This paper presented ITMER, a unified multimodal framework for emotion recognition integrating textual and visual modalities. Through systematic evaluation of text-only, image-only, and multimodal configurations, the study demonstrated that multimodal fusion significantly improves classification accuracy and macro-F1 performance.

Experimental results showed that multimodal models achieved up to 92% accuracy, substantially outperforming unimodal baselines. The findings confirm that emotional understanding benefits from the complementary integration of semantic and visual cues. Furthermore, the incorporation of Explainable AI techniques, including Grad-CAM and LIME, enhances model transparency by providing qualitative insight into both visual attention regions and influential textual tokens.

Overall, this work contributes A systematic unimodal–multimodal comparative evaluation, a robust image–text fusion framework, and integrated interpretability mechanisms for emotion recognition. Future research will explore attention-based fusion strategies, real-world paired datasets, and efficiency optimization for deployment in real-time human–computer interaction systems.

Code Availability

The code for all models and additional resources supporting the findings of this study is available at GitHub Repository.

Acknowledgements

The work was done with partial support from grants 20260626 (G.S.), 20260643 (A.G.), 20260367 (I.B.), and 20260496 (O.K.) by the Secretary of Research and Postgraduate Studies (SIP) of Instituto Politécnico Nacional, Mexico.

Author Contributions

F. and T. designed the study, conducted all experiments and developed the methodology. F.P.C. contributed to the original draft preparation. A.A. was responsible for writing, reviewing, and editing. R.Q.T. contributed to project administration, visualization, and supervision. All authors have read and agreed to the published version of the manuscript.

References

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artif Intell Rev*, 54(8):5789–5829.

Ameer, I. et al. (2023). Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Cai, L., Song, Y., Liu, T., and Zhang, K. (2020). A hybrid bert model that incorporates label semantics via adjustive attention for multi-label text classification. *IEEE Access*, 8:152183–152192.

Chandraumakantham, O., Gowtham, N., Zakariah, M., and Almazyad, A. (2024). Multimodal emotion recognition using feature fusion: an llm-based approach. *IEEE Access*, 12:108052–108071.

Chen, Y., Luo, H., Chen, J., and Wang, D. (2024). Multimodal emotion recognition algorithm based on graph attention network. In *Proceedings of the 5th international seminar on artificial intelligence, networking and information technology (AINIT)*, pages 814–822. IEEE.

Deshpande, U. U. et al. (2025). Multimodal sentiment analysis using image and text fusion for emotion detection. *Discover Computing*, 28(1):1–24.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. In

Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, pages 4171–4186.

Dong, Z., Li, X., Wang, Y., et al. (2022). Spontaneous facial expressions and micro-expressions coding: from brain to face. *Frontiers in Psychology*, 12:784834.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*.

Ezerceli, Ö. and Eskil, M. T. (2022). Convolutional neural network (cnn) algorithm based facial emotion recognition (fer) system for fer-2013 dataset. In *Proceedings of the international conference on electrical, computer, communications and mechatronics engineering (ICECCME)*, pages 1–6. IEEE.

Fei, Z., Yang, E., Li, D. U., et al. (2020). Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*, 388:212–227.

Feidakis, M., Daradoumis, T., and Caballé, S. (2011). Emotion measurement in intelligent tutoring systems: what, when and how to measure. In *Proceedings of the 3rd International Conference on Intelligent Networking and Collaborative Systems*, pages 807–812.

Geirhos, R. et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gong, L., He, X., and Yang, J. (2024). An image-text sentiment analysis method using multi-channel multi-modal joint learning. *Appl Artif Intell*, 38(1):2371712.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., et al. (2013). Challenges in representation learning: a report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer.

Hazmoune, S. and Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: taxonomies and state of the art review. *Eng Appl Artif Intell*, 133:108339.

Hosseini, S., Yamaghani, M. R., and Poorzaker Arabani, S. (2024a). A review of the methods of recognition multimodal emotions in sound, image and text. *International Journal of Applied Operational Research*, 12(1):29–41.

Hosseini, S. S., Yamaghani, M. R., and Poorzaker Arabani, S. (2024b). Multimodal modelling of human emotion using sound, image and text fusion. *Signal Image Video Process*, 18(1):71–79.

Jatoth, A., Abri, F., and Nguyen, T. (2025). Multimodal emotion detection and analysis from conversational data. *J Adv Inf Technol*, 16(8):1127.

Katada, S., Okada, S., and Komatani, K. (2022). Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*, 14(3):2443–2457.

Kement, U., Cavusoglu, M., Basar, B., and Kucunun, N. T. (2024). Facial emotion recognition research in the hospitality and tourism industry: a thematic content analysis. *Worldwide Hospitality and Tourism Themes*, 16(2):163–177.

Knyazev, B., Shvetsov, R., Efremova, N., and Kuharenko, A. (2018). Leveraging large face recognition data for emotion classification. In *Proceedings of the 13th IEEE international conference on automatic face and gesture recognition (FG 2018)*, pages 692–696. IEEE.

Kumar, P. et al. (2025). A multimodal deep learning framework for emotion recognition in text and visual media. In *Proceedings of the 8th International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 739–744.

Lan, Y.-T., Liu, W., and Lu, B.-L. (2020). Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Lapuschkin, S. et al. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.

Lian, Z. et al. (2025). Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint*.

Lian, Z., Sun, J., Wang, K., et al. (2024). Merbench: a unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint*.

Liang, Z., Chen, Y., Zhang, H., et al. (2024). A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Lorch, S., Gebele, J., and Brune, P. (2025). Towards trustworthy ai: evaluating shap and lime for facial emotion recognition. *arXiv preprint*.

Lu, W., Xu, Y., and Gu, J. (2025). Adaptive weighting in a transformer framework for multimodal emotion recognition. *Speech Commun*, page 103332.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops*, pages 94–101. IEEE.

Matsumoto, D. and Hwang, H. C. (2016). *APA Handbook of Nonverbal Communication*. American Psychological Association.

Nguyen, N. M., Nguyen, T. T., Tran, P. N., Lim, C. P., Pham, N. T., and Dang, D. N. M. (2026). Multimodal fusion in speech emotion recognition: A comprehensive review of methods and technologies. *Engineering Applications of Artificial Intelligence*, 163:112624.

Norval, M. and Wang, Z. (2025). Explainable artificial intelligence techniques for speech emotion recognition: a focus on xai models. *Inteligencia Artificial*, 28(76):85–123.

Panchal, V., Deshpande, T., Kandoi, T., Deulkar, K., and Narvekar, M. (2023). Multimodal emotion recognition in video, audio, and text using deep and transfer learning. In *Proceedings of the international conference on advanced computing technologies and applications (ICACTA)*, pages 1–7. IEEE.

Park, S. H., Bae, B. C., and Cheong, Y. G. (2020). Emotion recognition from text stories using an emotion embedding model. In *Proceedings of the IEEE international conference on big data and smart computing (BigComp)*, pages 579–583. IEEE.

Pereira, R. et al. (2024). Systematic review of emotion detection with computer vision and deep learning. *Sensors*, 24(11):3484.

Pons, G. and Masip, D. (2018). Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Sahin, E., Arslan, N. N., and Ozdemir, D. (2025). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 37(2):859–965.

Selvaraju, R. R. et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human–robot interaction: recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279.

Trinh, V. L. and Dao, T. L. (2022). Emotional speech recognition using deep neural networks. *Sensors*, 22(4):1414.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*.

Wang, Q., Wang, M., Yang, Y., and Zhang, X. (2022). Multi-modal emotion recognition using eeg and speech signals. *Computers in Biology and Medicine*, 149:105907.

Zhang, S., Wang, X., Liu, Y., et al. (2023). Multimodal emotion recognition based on audio and text by using hybrid attention networks. *Biomedical Signal Processing and Control*, 85:105052.

Zhao, J., Li, R., Jin, Q., Wang, X., and Li, H. (2022). Memobert: pre-training model with prompt-based learning for multimodal emotion recognition. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4703–4707. IEEE.