



# Multimodal Emotion Recognition for Emphatic Virtual Agents in Mental Health Interventions

Marcelo Alejandro Huerta-Espinoza, A.Y. Rodríguez González, Juan Martínez-Miranda  
Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic  
marcelo.huerta@cicese.edu.mx, ansel@cicese.mx, jmiranda@cicese.mx

Abstract Depression and anxiety disorders affect millions of individuals globally and are commonly addressed through psychological interventions. A growing technological approach to support such treatments involves the use of embodied conversational agents that employ motivational interviewing, a method that promotes behavioral change through empathic engagement. Despite its critical role in therapeutic efficacy, empathy remains a significant challenge for virtual agents to emulate. Emotion Recognition (ER) technologies offer a potential solution by enabling agents to perceive and respond appropriately to users' emotional states. Given the inherently multimodal nature of human emotion, unimodal ER approaches often fall short in accurately interpreting affective cues. In this work, we propose a multimodal emotion recognition model that integrates verbal and non-verbal signals (text and video) using a Cross-Modal Attention fusion strategy. Trained and evaluated on the IEMOCAP dataset, our approach leverages Ekman's taxonomy of basic emotions and demonstrates superior performance over unimodal baselines across key metrics such as accuracy and F1-score. By prioritizing text as the main modality and dynamically incorporating complementary visual cues, the model proves effective in complex emotion classification tasks. The proposed model is designed for integration into an existing conversational agent aimed at supporting individuals experiencing emotional and psychological distress. Future work will involve embedding the model in the conversational agent platform for emotionally distressed users, aiming to assess its real-world impact on engagement, user experience, and perceived empathy.

**Keywords:** Emotion recognition in conversation, Text, Image, Deep learning, Multimodal classification, Cross-Modal fusion.

## 1 Introduction

According to the World Health Organization (WHO), approximately 280 million people worldwide suffer from depression, representing around 3.8% of the global population [13]. Anxiety disorders affect an estimated 301 million people globally [1]. Both conditions are commonly treated with psychological interventions [13, 1]. One technological approach that complements such interventions is motivational interviewing with embodied conversational agents [23, 3]. When delivered by a virtual agent instead of a human, this approach integrates brief interventions with motivational strategies to promote healthier behavior change—an important factor in supporting mental and emotional health [23, 31].

A key principle of motivational interventions is empathy, emphasized by various authors as critical to both expressing and responding with understanding [32, 15, 3]. However, virtual agents still struggle to exhibit empathy, despite its importance in effective psychotherapeutic interventions and emotional support [12]. Empathy can be emulated through Emotion Recognition (ER), enabling virtual agents to respond appropriately to the user's emotional state.

While facial expressions are widely used as indicators of emotion [26, 22], verbal cues also carry rich emotional and semantic information. Therefore, combining verbal and non-verbal modalities enhances emotional interpretation. Unimodal approaches often fall short due to the complex, multimodal nature of human emotions [37]. In contrast, multimodal ER models integrate complementary signals, offering greater accuracy and resilience [4].

We propose a multimodal emotion recognition model for integration into an existing embodied conversational agent designed to support patients experiencing emotional and psychological distress [21]. The model leverages a Cross-Modal Attention fusion mechanism to integrate visual and textual cues. This modality selection enhances usability in noisy environments where speech input may be unreliable, and aligns with prior findings indicating users' preference for non-audio interaction [21]. The remainder of this paper is organized as follows: we first review related work on unimodal and multimodal ER; we then describe the feature extraction methods for each modality; next, we present our proposed model and its Cross-Modal Attention-based fusion strategy. The results and discussion are presented in Section 4; while Section 5 describes the conclusions and the future work.

## 2 Related Work

### 2.1 Image-Based Methods

Emotion recognition from images exploits the rich emotional content present in facial expressions. This task is commonly approached using deep learning techniques, particularly convolutional neural networks (CNNs) with pretrained models such as VGGNet and AlexNet [9, 6]. A notable example is the work by [14], which applied a CNN to the FER2013 dataset available on Kaggle. Another approach is used when video data is available, both spatial and temporal features can be captured. For example, [24] used 3D CNNs on the RAVDESS [19] and CREMA-D [7] datasets, while [38] combined CNNs with LSTM networks to model temporal dynamics in image sequences.

Although these methods achieve strong performance compared to other approaches, their reliance on unimodal datasets limits their ability to capture facial expressions in natural conversational contexts. This poses a significant challenge for conversational virtual agents and also results in the loss of the user's semantic content.

### 2.2 Text-Based Methods

Similarly, deep learning has dominated research efforts in text-based emotion recognition [2], particularly through the use of pretrained models such as BERT and RoBERTa. In [36], the performance of BERT and RoBERTa models was explored for dialogue data from the IEMOCAP dataset as a means of feature extraction. Likewise, [20] utilized BERT for feature extraction on the MELD and CMU-MOSEI datasets. These studies also highlight the relevance of these models for multimodal architectures, where after feature extraction, the representations are fused into multimodal models, as is also the case in [16], where BERT was employed on the IEMOCAP dataset.

While text-based methods have advanced emotion recognition, their unimodal focus overlooks non-verbal cues, leading to a partial understanding of affective states. Additionally the frequent omission of Ekman's emotion taxonomy reduces applicability in universally oriented emotional models

### 2.3 Fusion Data Methods

Among the most extensively studied approaches in multimodal emotion recognition are the models proposed in [34], HU-Dialogue [8] and Multilogue [30]. The models demonstrates strong performance through inter-modal fusion with high computational robustness using LSTM and GRU architectures, while the latter utilizes GRUs for each modality combined with a pairwise cross-modal attention mechanism that captures interactions between modality representations and conversation participants. Other state-of-the-art models include Transmodality [33], a multimodal fusion model based on transformer architectures, and IMAN [25], which performs fusion using an attention-based mechanism.

Recent trends favor attention-based fusion models, such as those in [39] and [16], which use cross-modal Transformers for fusing speech, text, and visual modalities, showing strong results on datasets like IEMOCAP. However, these approaches also omit Ekman’s emotion set and rely on computationally expensive Transformer-based components, making them less practical for interactive systems.

### 3 Proposed Method

This section outlines the methodological framework adopted in this study. In order to validate the improvement introduced by the proposed fusion model over unimodal models, separate unimodal architectures were first developed for visual and textual data. Subsequently, the same architectural structure and pretrained weights were employed in the development of the fusion model.

#### 3.1 Data

The IEMOCAP [5] dataset was selected for validate our proposed method due to its widespread use in the field of multimodal emotion recognition and its specific design for capturing emotional expressions within conversational contexts. IEMOCAP comprises video recordings and corresponding transcriptions, both annotated with emotional labels. This dataset includes a total of nine annotated emotions, as shown in Table 1.

The emotional taxonomy adopted in this study is based on Ekman’s model of basic emotions [11], which defines six universal emotional categories: anger, happiness, neutral, sadness, surprise, fear, and disgust. These emotions were selected because they are considered representative of the most commonly expressed affective states in human-agent interaction scenarios. To the best of our knowledge, there is a lack of studies applying Ekman’s taxonomy directly to the IEMOCAP dataset, which motivated the inclusion of both Ekman’s six emotions (Emotion Set 1) and the six most prevalent emotions in IEMOCAP (anger, happiness, neutral, sadness, frustration, and excitement, which we will know as Emotion Set 2) in our experiments. Emotion Set 1 will be used within the empathic conversational agent, and it will not be compared with other state-of-the-art models due to the lack of existing studies using this specific set of emotions. In contrast, Emotion Set 2 will be employed to evaluate the fusion model architecture against other state-of-the-art approaches. Table 1 indicates which emotions are included in each of the two groups considered in this study. This dual approach enables comparison with prior studies while ensuring consistency across unimodal and multimodal model architectures, which were maintained as similar as possible.

Table 1: Distribution of emotion labels in the IEMOCAP dataset.

<b>Emotion</b>	<b>Number of Instances</b>	<b>Emotion Set</b>
Frustration	1849	Emotion Set 2
Neutral	1708	Both
Anger	1103	Both
Sadness	1084	Both
Excitement	1041	Emotion Set 2
Happiness	595	Both
Surprise	107	Emotion Set 1
Fear	40	Emotion Set 1
Disgust	2	Emotion Set 1

#### 3.2 Data Processing

This section describes the two types of data used (image and text). These channels were selected because the agent for which this model is designed relies on textual and visual modalities to interact with users [21]. Furthermore, during testing, users reported that they were not comfortable speaking directly to

the agent, which can be deployed as either a mobile or desktop application. They preferred to use it in environments where there might be significant background noise. For this reason, the agent does not utilize the audio channel, this is the main reason of using text and visual data.

### 3.2.1 Image Data Processing

The video recordings in the IEMOCAP dataset were captured at a frame rate of 30 frames per second. As a first step, individual frames were extracted at the same sampling rate and stored according to their corresponding utterance and emotion label. In the subsequent preprocessing phase, only facial regions were retained for analysis. To achieve this, the MTCNN face detector [35], accessed via the DeepFace framework [29, 28], was employed. This preprocessing step aimed to eliminate background noise and non-relevant visual information, thus enhancing the model’s capacity to learn emotionally salient facial features.

### 3.2.2 Text Data Processing

The textual modality required comparatively less preprocessing, as transfer learning techniques were utilized through pretrained transformer-based language models. For experiments using Ekman’s taxonomy, the BERT-base model (Bidirectional Encoder Representations from Transformers) [10] was selected due to its robust performance across diverse natural language processing tasks, including emotion recognition. In contrast, for experiments aligned with the standard IEMOCAP emotion labels, the RoBERTa-base model [18], a retrained and optimized version of BERT, was used. Both models incorporate their own tokenizers, which convert textual input into numerical embeddings tailored to classification tasks. The only preprocessing applied to both unimodal models consisted of tokenization using the respective pretrained tokenizers.

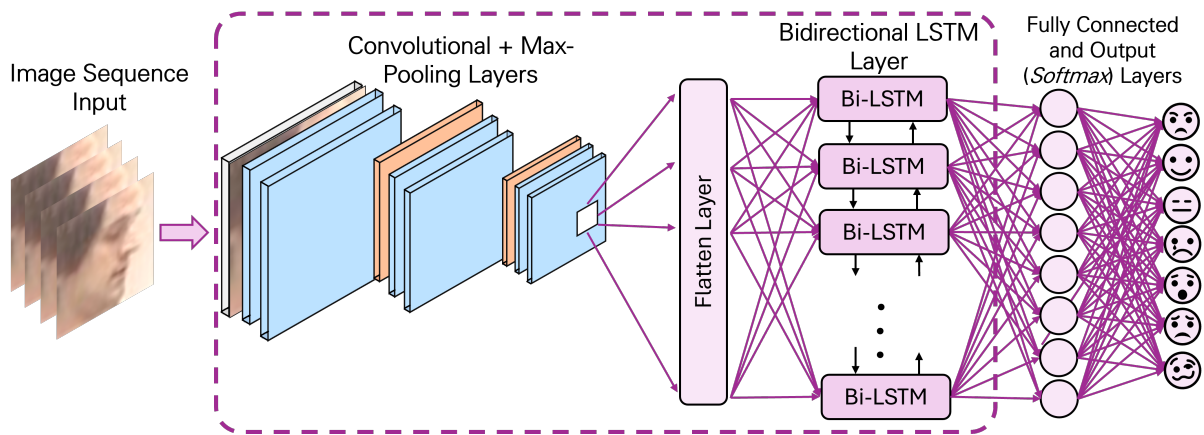


Figure 1: Unimodal image model.

## 3.3 Unimodal Feature Extraction

This section explains the process employed in the model design and feature extraction process in the unimodal models.

### 3.3.1 Image Feature Extraction

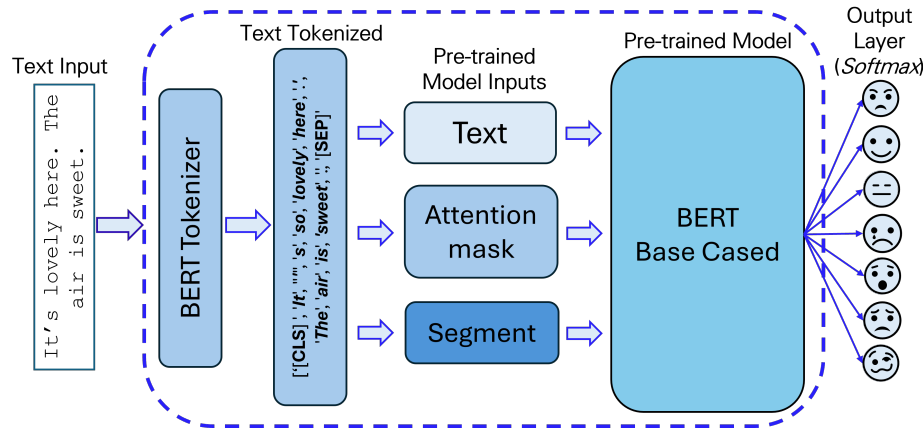
For both emotion sets (Emotion Set 1 and Emotion Set 2), a feature extraction approach inspired by several state-of-the-art studies was employed. This method involves a two-stage process: first, the extraction of spatial features from image vectors, followed by the modeling of their temporal dynamics. Figure 1 illustrates the architecture used to address these two tasks. Specifically, spatial features were

obtained using a convolutional neural network composed of three convolutional layers, each followed by a Max Pooling operation to progressively reduce spatial dimensionality while preserving the most salient information.

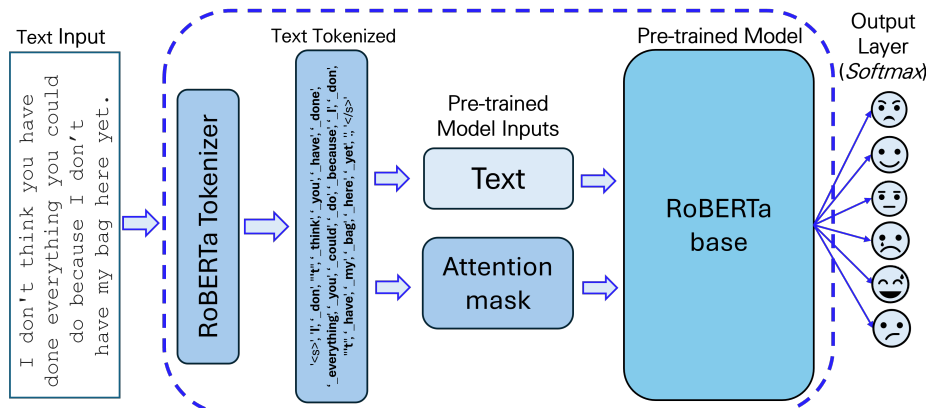
To capture the temporal evolution of facial expressions, the sequence of convolutional features is fed into a bidirectional Long Short-Term Memory (Bi-LSTM) network. This recurrent structure models both forward and backward dependencies, which is crucial in emotion recognition where the order and duration of expressions convey meaningful emotional states. The resulting feature vector is projected into a latent space via a dense layer with activation function, followed by a softmax layer to compute the probability distribution over the emotion classes.

### 3.3.2 Text Feature Extraction

As discussed in previous sections, the unimodal text models were based on the pretrained BERT [10] and RoBERTa [18] architectures. Figure 2 illustrates the overall architecture employed for both sets (Emotion Set 1 and 2). As shown, the processing pipeline is largely similar for both models, with the main differences residing in the choice of pretrained model and tokenizer.



(a) Unimodal text model Emotion Set 1.



(b) Unimodal text model Emotion Set 2.

Figure 2: Unimodal text architectures.

Notably, as depicted in Figure 2a and 2b, the RoBERTa tokenizer does not utilize segment embeddings, which distinguishes it from BERT's tokenization scheme. In both cases, the output of the transformer-based encoder is passed to a softmax classification layer, which produces the final probability distribution over the emotion classes.

### 3.4 Cross-Modal Fusion Strategy

The fusion model, as described in previous sections, integrates the two unimodal architectures (excluding their respective dense and classification layers) and incorporates a Cross-Modal Attention mechanism. This mechanism facilitates the interaction between modalities by enabling the model to dynamically align and integrate complementary information from each source. The fusion architecture is composed of three fundamental components:

- **Query (Q):** Represents the modality being queried or attended to; it defines the focus of the attention mechanism.
- **Key (K):** Encodes the modality that provides contextual cues to guide the attention over the query.
- **Value (V):** Contains the content from which information is extracted, modulated by the attention weights.

Using these three components, the Cross-Modal Attention mechanism learns an attention weight matrix ( $W_{\text{Cross-Modal}}$ ) that captures relevant intermodal interactions during training to enhance classification performance. Several experiments were conducted by varying the assignment of modalities to the  $Q$ ,  $K$ , and  $V$  components. The best-performing configuration used the text modality as the primary input (assigned to  $Q$ ), while the visual modality provided contextual information (assigned to both  $K$  and  $V$ ). This configuration leveraged a multi-head attention mechanism with four attention heads.

To formally establish text as the primary modality, it was designated as the query ( $Q = F_{\text{txt}}$ ), while the visual features were assigned as key and value ( $K = V = F_{\text{img}}$ ), where  $F_{\text{txt}}$  and  $F_{\text{img}}$  represent the feature tensors extracted from the text and image modalities, respectively. The overall fusion architecture is illustrated in Figure 3. The final output of the Cross-Modal Attention mechanism is computed according to Equation 1:

$$\hat{y} = \text{softmax}(W_{\text{Cross-Modal}}(F_{\text{img}}, F_{\text{txt}})) \tag{1}$$

where  $\hat{y}$  denotes the predicted emotion label produced by the model.

Following the Cross-Modal Attention mechanism, the resulting fused representation is concatenated with the textual features. This combined feature vector is then flattened and passed through a dense classification layer to generate the final output.

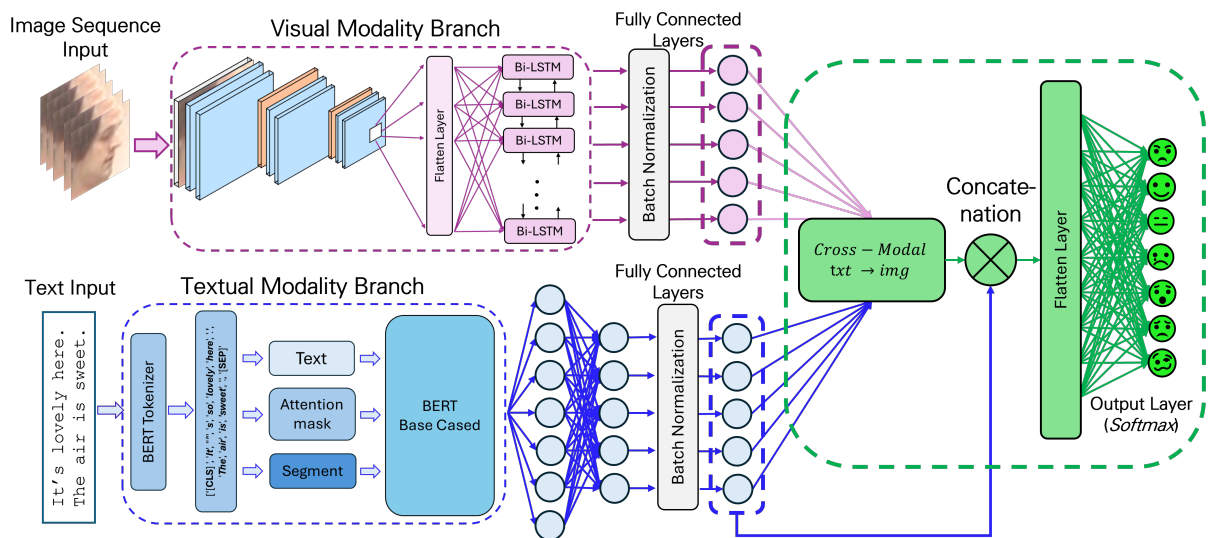


Figure 3: Cross-Modal fusion architecture. The fusion module is highlighted in green.

### 3.5 Training Details

#### 3.5.1 Hyperparameters Settings

The unimodal architecture for processing image sequences consisted of three convolutional layers with 32, 64, and 128 filters, respectively. Each convolutional layer was followed by a Max Pooling operation with a pooling size of 2, allowing progressive spatial feature reduction. To model temporal dependencies in the sequence of facial features, a bidirectional Long Short-Term Memory (BiLSTM) layer was employed. A dropout rate of 0.25 was applied to mitigate overfitting. The output was then passed through a dense layer with 64 neurons. Finally, a softmax classification layer was added, with the number of neurons corresponding to the number of emotion categories: 7 for Emotion Set 1 and 6 for Emotion Set 2.

For the text modality, the architecture was based on pretrained BERT or RoBERTa models, as previously described. These models were directly followed by a softmax classification layer configured with the same number of output neurons as in the image model, depending on the target emotion group.

Before applying the Cross-Modal Attention mechanism, the textual feature representations were passed through a dense layer to match the dimensionality of the image sequence features. Both the textual and visual features were then normalized using Batch Normalization to ensure consistency across modalities.

The fusion was carried out using a Cross-Modal Attention mechanism, where the text modality was designated as the query and attended to the image modality, serving as both key and value. A multi-head attention mechanism with 4 heads was used to enhance the model's capacity to capture intermodal dependencies. The resulting fused features were flattened and passed through a final dense classification layer, again adapted to each Emotion Set with 7 or 6 output units. Additionally all models were trained using a batch size of 8 and were trained for a total of 20 epochs.

#### 3.5.2 Loss Function and Optimizer

The proposed model was trained using the cross-entropy loss function, which is widely adopted in multi-class classification problems. This loss quantifies the divergence between the ground-truth class distribution  $y$  (expressed in one-hot encoding) and the predicted distribution  $\hat{y}$ . Formally, it is defined as:

$$\mathcal{L}(y, \hat{y}) = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (2)$$

where  $C$  denotes the total number of classes,  $y_i$  indicates the true class label (equal to 1 if the ground-truth class corresponds to  $i$ , and 0 otherwise), and  $\hat{y}_i$  represents the predicted probability for class  $i$ , obtained by Equation 1. Model parameter optimization was carried out using the AdamW optimizer [17], a variant of Adam that decouples the weight decay term from the gradient-based update associated with adaptive moment estimation. The optimizer was configured with a learning rate of  $\theta = 0.001$ , which facilitated stable and efficient convergence throughout the training phase.

#### 3.5.3 Evaluation

The model performance was assessed using two standard classification metrics: accuracy and F1-score. Accuracy was utilized as the primary metric to measure the overall classification performance. The F1-score was used to account for class imbalance by combining both precision and recall into a single harmonic mean.

## 4 Results and Discussion

As discussed in Section 3.1, to the best of our knowledge, there is a lack of studies that apply Ekman's taxonomy directly to the IEMOCAP dataset. Therefore, we only compare Emotion Set 2 against other state-of-the-art models. To evaluate the performance of the proposed multimodal emotion recognition model, both quantitative metrics (Accuracy and F1-score) and qualitative analyses through confusion matrices were examined for both emotion sets.

Table 2: Performance by modality for Emotion Set 1

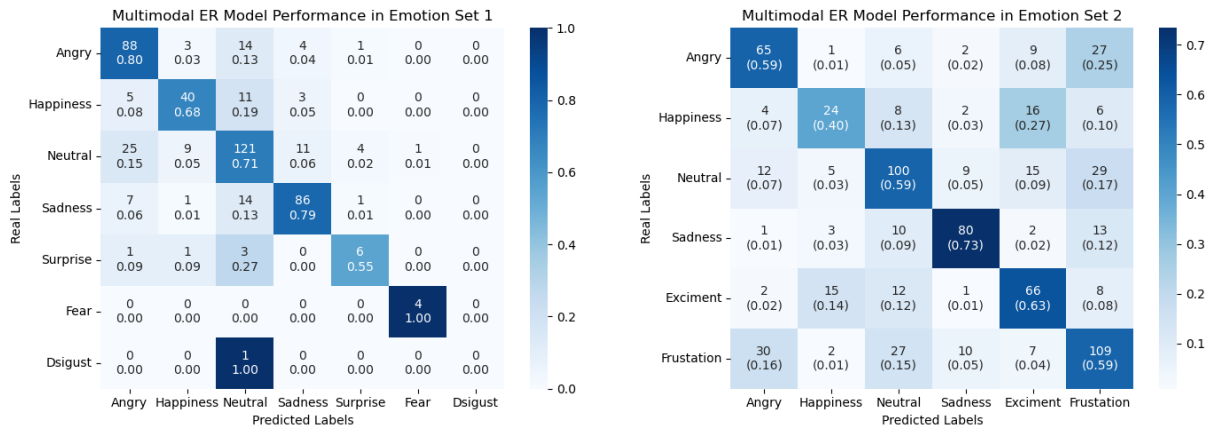
	Emotion Set 1		Emotion Set 2	
	Accuracy	F1-Score	Accuracy	F1-Score
Only Visual data	0.59	0.33	0.46	0.46
Only Text data	0.67	0.59	0.54	0.53
Visual + Text (Text as the main modal)	<b>0.74</b>	<b>0.63</b>	<b>0.60</b>	<b>0.59</b>

As shown in Table 2, the model that leverages both visual and textual data—prioritizing text as the primary modality—consistently outperforms unimodal approaches across both emotion sets. In Emotion Set 1, the multimodal configuration achieved the highest accuracy (0.74) and F1-score (0.63), compared to the text-only model (accuracy = 0.67, F1 = 0.59) and the visual-only model (accuracy = 0.59, F1 = 0.33). A similar trend was observed in Emotion Set 2, albeit with lower overall performance: the multimodal setup attained an accuracy of 0.60 and an F1-score of 0.59, surpassing both text-only (accuracy = 0.54, F1 = 0.53) and visual-only models (accuracy = 0.46, F1 = 0.46).

These results underscore the effectiveness of multimodal integration in emotion recognition tasks, particularly highlighting the critical role of textual information in capturing affective cues in complex interactions.

The confusion matrices depicted in Figure 4 provide further insight into class-wise model behavior:

- For Emotion Set 1, the model exhibits strong classification performance for emotions such as *Sadness* (F1  $\approx$  0.79), *Neutral* (F1  $\approx$  0.71), and *Angry* (F1  $\approx$  0.80). Conversely, classes like *Happiness* and *Surprise* show greater confusion with neighboring emotions, indicating challenges in distinguishing affectively similar states.
- Emotion Set 2 presents more variability in classification accuracy, with a general decline in per-class performance. Notably, *Angry* is frequently misclassified as *Frustration* (27 instances), suggesting either perceptual overlap in expression or data distribution challenges. Despite this, the model still performs reasonably well on emotions such as *Sadness* (F1  $\approx$  0.73) and *Frustration* (F1  $\approx$  0.59).



(a) Confusion matrix of the multimodal model (Emotion Set 1).

(b) Confusion matrix of the multimodal model (Emotion Set 2).

Figure 4: Multimodal Models performance

The decrease in performance observed in Emotion Set 2 may reflect increased individual variability in emotional expression, differing data quality, or context-specific ambiguities. Table 3 presents a comparative analysis between our multimodal model and other state-of-the-art approaches on Emotion Set

Table 3: Performance of different models against our in Emotion Set 2

Model	Modalities	Accuracy	F1-Score
Transmodality [33]	Speech + Visual + Text	0.608	-
A-DMN [34]	Speech + Visual + Text	0.646	0.643
IMAN [25]	Speech + Visual + Text	0.650	0.645
HU-Dialogue [8]	Speech + Visual + Text	0.657	0.653
<b>Ours</b>	Visual + Text	0.601	0.591

2. As discussed in Section 3.1, there is a lack of studies applying Ekman’s taxonomy; therefore, we limit our comparison of results to Emotion Set 2.

While our model, which integrates only visual and textual modalities, reports slightly lower performance (accuracy = 0.601, F1-score = 0.591) compared to more complex models such as HU-Dialogue (accuracy = 0.657, F1 = 0.653) and IMAN (accuracy = 0.650, F1 = 0.645), it remains competitive despite the absence of the speech modality. This performance gap is expected, given that models like A-DMN, IMAN, and HU-Dialogue leverage speech data in addition to visual and textual cues-offering a more comprehensive multimodal context. However, it is important to emphasize that the exclusion of audio in our approach was not arbitrary. As discussed in Section 1, our model was designed to align with the capabilities and constraints of the target application: a virtual agent platform that processes only text and visual input from users. Thus, the choice of modalities is application-driven and directly reflects real-world deployment scenarios. Furthermore, by achieving performance comparable to that of audio-enhanced models, our approach highlights the potential of lightweight, audio-free architectures in scenarios where speech is unavailable, such as text-based virtual assistants or privacy-sensitive environments.

## 5 Conclusion and Future Work

This study presents a multimodal emotion recognition model that integrates visual and textual information, for a virtual conversational agent providing psychotherapeutic-based sessions. Through comprehensive evaluation across two user groups, our findings demonstrate that the proposed model effectively leverages both modalities to outperform unimodal baselines, achieving its best performance in Emotion Set 1 with an accuracy of 0.74 and F1-score of 0.63. Even in the more challenging Emotion Set 2, the model maintains competitive performance, despite increased variability in emotional expression.

The proposed model offers a practical and efficient solution for emotion recognition in scenarios where audio input is unavailable or undesirable, thereby broadening the accessibility and applicability of affective computing technologies. These capabilities enable its seamless integration into applications where accurate emotion recognition is essential to provide effective psychological interventions for users.

Notably, when benchmarked against state-of-the-art models that incorporate speech, our approach remains robust and competitive. While some performance gap exists, this is expected given the richer multimodal input used by those models. Crucially, our design decisions were guided by the intended application context-namely, a virtual agent that interacts with users via visual and textual modalities only. In this sense, the exclusion of audio is not a limitation but a deliberate adaptation to real-world constraints.

As future work, we plan to integrate the multimodal model into the conversational virtual agent and conduct user studies to evaluate its perceived empathic behavior. We intend to design an emotionally charged scenario to elicit emotions outside of a strictly mental health context. The goal is to first assess whether the agent can accurately detect emotions and respond empathetically to users’ comments. We also plan to compare the performance of the conversational agent with access to emotional information against a version without such information. It is expected that higher education students will participate in the study. In order to assess the system’s perceived empathy, a validated and state-of-the-art instrument has been selected. Specifically, the PETS scale [27] will be employed, a 10-item, 2-factor instrument designed to rigorously measure and compare perceived empathy in interactive systems.

The authors gratefully acknowledge the financial support provided to the first author by SECIHTI

which supported his graduate studies during the development of this work.

## References

- [1] Jordi Alonso, Zhaorui Liu, Sara Evans-Lacko, Ekaterina Sadikova, Nancy Sampson, Somnath Chatterji, Jibril Abdulmalik, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Laura H Andrade, et al. Treatment gap for anxiety disorders is global: Results of the world mental health surveys in 21 countries. *Depression and anxiety*, 35(3):195–208, 2018.
- [2] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, 2020.
- [3] Hal Arkowitz, William R. Miller, Henny A. Westra, and Stephen Rollnick. Motivational interviewing in the treatment of psychological problems: Conclusions and future directions. In *Motivational Interviewing in the Treatment of Psychological Problems*, Applications of Motivational Interviewing, pages 324–342. The Guilford Press, New York, NY, US, 2008.
- [4] Geetha A.v., Mala T., Priyanka D., and Uma E. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218, May 2024.
- [5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.
- [6] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.
- [7] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.
- [8] Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. *IEEE Transactions on Cybernetics*, 54(1):187–198, 2024.
- [9] Andrada-Livia Cirneanu, Dan Popescu, and Dragos Iordache. New trends in emotion recognition using image analysis by neural networks, a systematic review. *Sensors*, 23(16):7092, 2023.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.
- [11] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.
- [12] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. Empathy. *Psychotherapy*, 48(1):43, 2011.
- [13] SAGS Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, Wai-Tat Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the who world mental health (wmh) surveys. *Psychological medicine*, 48(9):1560–1571, 2018.
- [14] Yijun Gan. Facial expression recognition using convolutional neural network. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, ICVISIP 2018, New York, NY, USA, 2018. Association for Computing Machinery.

- [15] Kate Hall, Tania Gibbie, and Dan I Lubman. Motivational interviewing techniques: Facilitating behaviour change in the general practice setting. *Australian family physician*, 41(9):660–667, 2012.
- [16] Mustaqeem Khan, Phuong-Nam Tran, Nhat Truong Pham, Abdulmotaleb El Saddik, and Alice Othmani. Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*, 15(1):5473, 2025.
- [17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [19] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS one*, 13(5):e0196391, 2018.
- [20] Fazliddin Makhmudov, Alpamis Kultimuratov, and Young-Im Cho. Enhancing multimodal emotion recognition through attention mechanisms in bert and cnn architectures. *Applied Sciences*, 14(10):4199, 2024.
- [21] Juan Martínez-Miranda, Ariadna Martínez, Roberto Ramos, Héctor Aguilar, Liliana Jiménez, Hodwar Arias, Giovanni Rosales, and Elizabeth Valencia. Assessment of users’ acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour. *Journal of medical systems*, 43(8):246, 2019.
- [22] Albert Mehrabian. Communication without words. In *Communication theory*, pages 193–200. Routledge, 2017.
- [23] José Mercado, Ismael Edrein Espinosa-Curiel, and Juan Martínez-Miranda. Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *Journal of medical Internet research*, 25:e52097, 2023.
- [24] Bogdan Mocanu, Ruxandra Tapu, and Titus Zaharia. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing*, 133:104676, 2023.
- [25] Minjie Ren, Xiangdong Huang, Xiaoqi Shi, and Weizhi Nie. Interactive multimodal attention network for emotion recognition in conversation. *IEEE Signal Processing Letters*, 28:1046–1050, 2021.
- [26] Nastaran Saffaryazdi, Tamil Selvan Gunasekaran, Kate Laveys, Elizabeth Broadbent, and Mark Billingham. Empathetic conversational agents: Utilizing neural and physiological signals for enhanced empathetic interactions. *arXiv preprint arXiv:2501.08393*, 2025.
- [27] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. Perceived empathy of technology scale (pets): Measuring empathy of systems toward the user. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery.
- [28] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024.
- [29] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.
- [30] Aman Shenoy and Ashish Sardana. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria, editors, *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA, July 2020. Association for Computational Linguistics.

- 
- [31] Andrew Steptoe. Health behavior and stress. In George Fink, editor, *Encyclopedia of Stress*, pages 262–266. Academic Press, New York, NY, 2nd edition, 2007.
- [32] Janet Treasure. Motivational interviewing. *Advances in Psychiatric Treatment*, 10(5):331–337, 2004.
- [33] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of the web conference 2020*, pages 2514–2520, 2020.
- [34] Songlong Xing, Sijie Mai, and Haifeng Hu. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3):1426–1439, 2020.
- [35] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016.
- [36] Shaohua Zhang, Yan Feng, Yihao Ren, Zefei Guo, Renjie Yu, Ruobing Li, and Peiran Xing. Multimodal emotion recognition based on wavelet transform and bert-roberta: An innovative approach combining enhanced bilstm and focus loss function. *Electronics*, 13(16):3262, 2024.
- [37] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692, March 2024.
- [38] ShiHao Zou, Xianying Huang, XuDong Shen, and Hankai Liu. Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. *Knowledge-Based Systems*, 258:109978, December 2022.
- [39] ShiHao Zou, Xianying Huang, XuDong Shen, and Hankai Liu. Improving multimodal fusion with main modal transformer for emotion recognition in conversation. *Knowledge-Based Systems*, 258:109978, 2022.