Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 13 July 2026
Revised 23 Aug 2026
Accepted 11 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (20s): 1103–1114

KEYWORDS:
3D CNN-LSTM, BERT, CMU-MOSEI, Cross-attention transformer, Emotion intensity, MELD, Multimodal emotion recognition, MUStARD++, ResNet-50, Supervised contrastive learning

MEXA: A Multi-Modal Emotion eXtraction and Analysis Framework with Cross-Attention Fusion and Contrastive Sarcasm Detection

Bindu K[1], Dr. Manjunatha R C[2]

[1]Assistant Professor, Department of ECE, Global Academy of Technology, Bengaluru.
[2]Professor, Department of AIML, M S Engineering College, Bengaluru.

ABSTRACT: In digital communication, emotion analysis needs to be multimodal, that is, the integration of heterogeneous modalities such as language, facial expression and temporal visual dynamics. Most of the existing methods, however, are mainly unimodal and fail to consider the inconsistencies that are common between modalities which are crucial for sarcasm recognition. To overcome this drawback, this paper proposes MEXA (Multimodal Emotion eXtraction and Analysis), which is a three-branch deep learning framework for a unified categorical emotion recognition, emotion-intensity regression on a [0, 3] scale, and binary sarcasm detection. MEXA generates textual representations, visual features and temporal visual representations, by applying a fine-tuned 12-layer BERT model, a ResNet-50 backbone pretrained on AffectNet and a 3D CNN-LSTM model, respectively. A cross-attention Transformer fusion module is then used to fuse the information from the three modalities, enabling each to attend to complementary representations from the other modalities. To boost the discrimination of multimodal representations, MEXA also adds a supervised InfoNCE-based contrastive learning objective for sarcasm detection. The framework is tested and assessed on three metrics: MUStARD++ (Sarcasm Detection), CMU-MOSEI (Emotion-Intensity Estimation), and MELD (Emotion Recognition in Conversational Environments). With a weighted F1-score of 75.8%, MEXA is 1.3 percentage points ahead of the best reported results on MUStARD++. It has an average MAE of 0.759 on CMU-MOSEI for emotion-intensity regression, and a weighted F1-score of 65.4% on MELD for emotion recognition in seven classes. Due to substantial class imbalance in MELD (approx. 47% of samples are in the Neutral class, while Fear and Disgust make up about 2% of samples each), weighted F1 is used for the evaluation throughout.

1. INTRODUCTION

The detection of emotion in digital communication is a specialised form of NLP that is more than just sentiment polarity, it identifies feelings like joy, anger, fear, sadness, disgust and surprise [1]. Emotion capture in multi-modal modalities (text, face imagery, video) is difficult, as human expression is inherently multimodal: a smile on the face contradicting the verbal compliment indicates sarcasm; the rhythmic flow of the visual information across the video frames indicates stress or anxiety. This work is motivated by three recurring limitations: (i) Unimodal dominance: Majority of systems are mainly text-based with other modalities treated as auxiliary [2, 3]; (ii) Temporal blindness: Static image encoders do not perform well on micro-expressions that are captured across consecutive frames [10, 11]; (iii) Sarcasm-agnostic training: Emotion classifiers trained on sincere expressions suffer from poor generalization to sarcastic utterances [12, 13].

porttitor, facilisis luctus, metus. Phasellus ultrices nulla quis nibh. Quisque a lectus. Donec consectetuer ligula vulputate sem tristique cursus.

To overcome the three limitations of previous approaches, this paper introduces MEXA, a multimodal approach. and (C3) a multi-task teacher that fine-tunes the three branch encoders, resulting in 3D CNN-LSTM, BERT, and ResNet-50 models pre-

trained on the 3D Motion and Expression dataset that are capable of extracting features from images, videos, and text, respectively. The main advantages are the following: (C1) a three-branch encoder that independently extracts textual, visual, and temporal features from the data using BERT, ResNet-50, and 3D CNN-LSTM, respectively; (C2) a cross-attention Transformer-based fusion module for enabling information exchange between the 3D branches; and (C3) a multi-task teacher that fine-tunes the three branches of the encoder to learn features from 3D images, videos, and text, respectively, using the 3D Motion and Expression dataset,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

A supervised InfoNCE-based contrastive learning objective to enhance multimodal representations for sarcasm detection, (C4) a joint multi-task learning objective,

$$L_{\text{total}} = L_{\text{cls}} + \alpha L_{\text{reg}} + \beta L_{\text{sarc}} + \gamma L_{\text{irony}} \qquad (2)$$

and (C5) a full mathematical model of the proposed framework in Section IV that will be helpful in reproduction. The proposed model is competitive and realistic on three benchmark datasets. On MUStARD++, it gets a weighted F1 of 75.8%, which is 1.3 percentage point higher than reported state of the art. On CMU-MOSEI, it has an average MAE of 0.759 for emotion-intensity prediction on the 03scale. MELD results show that the model achieves a weighted F1 score of 65.4% in seven-class emotion recognition, and the macro-F1 score of ~48.2% indicates the influence of class imbalance.

The paper is organized as follows: Section II outlines the related works, Section III presents the architecture and the proposed methodology of the system, Section IV describes the datasets used for the paper, Section V reports the experimental results along with the ablation studies and state-of-the-art comparison. Finally, Section VI concludes the paper.

2. MATERIALS AND METHODS

2.1 Study Material

A. Multimodal Emotion Recognition

Sun et al. [7] (ICASSP 2023) fused Wav2Vec 2.0 and BERT with auxiliary tasks for speech-text emotion, getting CMU-MOSEI Acc-7 of 51.6%. Sun et al. [8] (IEEE TAFFC 2023) presented efficient multimodal transformer models with dual-level feature restoration with MAE 0.812 on CMU-MOSEI emotion intensity. To overcome this, TACFN [9] proposed adaptive cross-modal fusion which showed improvements over static cross-attention. Cross-modal transformer feature fusion on speech-text pairs was employed in MemoCMT [14] (Scientific Reports 2025) to get CMU-MOSEI Acc-7 of 55.9%.

B. Sarcasm Detection on MUStARD/MUStARD++

The multimodal sarcasm detection benchmark was created by the original MUStARD dataset [5] which consists of 690 clips of TV shows. MUStARD++ [Ray et al., 2022] is a more complex evaluation scenario which is an extended version of the original MUStARD. The reported state-of-the-art performance on MUStARD++ is around 74.6% weighted F1, obtained by Shi's multimodal early-fusion approach [2024]. For text-image contrastive debiasing, Jia et al. [12] (AAAI 2024) got 74.2% weighted F1, and Wei et al. [13] (Knowledge-Based Systems, 2024) got 74.1% F1 with graph-based contrastive learning. By contrast, the results obtained for the original MUStARD data set are significantly higher as this is a relatively easier evaluation setting. The current state-of-the-art results are 81.64% macro-F1 for PC-MNet [2025] with all three modalities (text, audio, and video), and 81.0% text-only audio modality F1 score for AMuSeD [2024]. MEXA focuses on the more challenging MUStARD++ benchmark and obtains 75.8% weighted F1, which is a 1.3 percentage point improvement over the previous report of the best performance on this benchmark.

C.MELD Emotion Recognition - Metric Clarification

The MELD dataset [6] is also highly imbalanced, having about 47% of the total utterances in the Neutral category and about 2% in the Fear and Disgust categories. This means that this F1 (and the macro-F1) can vary significantly, with an average of 17-22 percentage points. The majority of previous work on MELD, such as M2FNet, MMGCN, AIMDiT, and DialogueCRN, report weighted F1-scores, and the performance reported ranges from 58% to 67%. On the other Hand, macro-F1 considers all the seven emotion classes equally important, regardless of how common they are, and will generally get lower scores if there is such

a disparity, which is usually in the range of 35–55%. This study adopts the convention of the previous research to report weighted F1 as the main metric of the study to ensure a fair comparison with earlier studies. Reported separately, Macro-F1 is a metric which is also added to give a more balanced assessment of the performance on the minority classes and to avoid potentially misleading conclusions from the large class imbalance in MELD.

MEXA is a jointly-trained multimodal emotion analysis system. Fig. 1 shows the architecture. Full mathematical derivations are in the upcoming subsections.

MEXA Architecture Overview

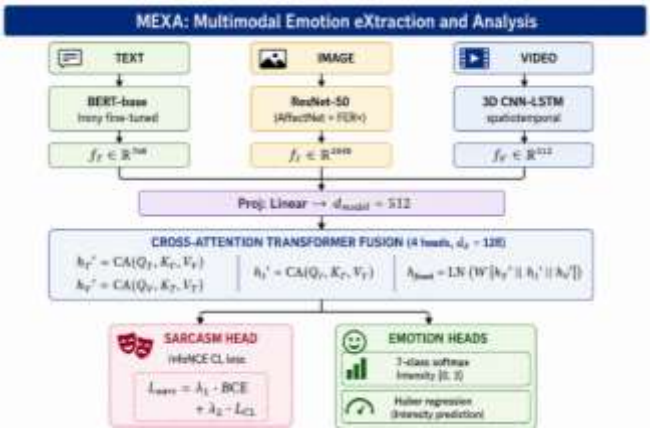


Fig. 1. MEXA Architecture: Three-Branch Encoder with Cross-Attention Fusion, Contrastive Sarcasm Head, and Emotion Heads

In the proposed architecture MEXA, the text, image and video modalities are combined with the help of BERT-base, ResNet-50 and a 3D CNN-LSTM. The extracted features are then mapped to a shared 512 dimensional space and combined using a cross-attention Transformer. A separate sarcasm head and emotion head can then be used to fuse the representation, which can be used to detect sarcasm and classify the emotion intensity and emotion class using contrastive learning, softmax, and Huber regression.

Problem Formulation

A multimodal utterance is defined as $X = (u_t, x_I, x_V)$, which consists of a textual utterance u_t, a face image $x_I \in \mathbb{R}^{H\times W\times 3}$ and a video clip of length $T = 16$ frames, $x_V \in \mathbb{R}^{T\times H_v\times W_v\times 3}$ with H_v and W_v denoting the height and width of the video clip. MEXA jointly performs three tasks: seven-class emotion classification, six-dimensional emotion-intensity regression on the original $[0, 3]$ scale, and binary sarcasm detection:

$$P(y_e \mid X) = \text{softmax}\big(f_e(h_{\text{fused}})\big) \in \mathbb{R}^7 \quad (3)$$

$$\hat{e} = \big[\hat{e}_{\text{joy}}, \dots, \hat{e}_{\text{surp}}\big] \in [0, 3]^6 \quad (4)$$

$$P(y_s = 1 \mid X) = \sigma(W_s h_{\text{fused}} + b_s) \quad (5)$$

BERT-Based Text Encoding

The input sequence is represented as $U = [\text{CLS}, w_1, \dots, w_n, \text{SEP}]$. Each token is encoded by combining token, positional, and segment embeddings:

$$e_i = E_{\text{tok}}(w_i) + E_{\text{pos}}(i) + E_{\text{seg}}(s_i) \quad (6)$$

The resulting representation is processed through 12 Transformer layers using multi-head self-attention. For layer l,

$$Q^{(l)} = H^{(l-1)}W_Q^{(l)},\ K^{(l)} = H^{(l-1)}W_K^{(l)},\ V^{(l)} = H^{(l-1)}W_V^{(l)} \quad (7)$$

For each attention head, $\text{head}_j = \text{softmax}\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right) V_j,\ d_k = 64$ (8)

with 12 attention heads. The outputs are concatenated and projected as:

$$\text{MHA}\left(H^{(l-1)}\right) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{12}) W_O^{(l)} \quad (9)$$

Residual connections and layer normalization are then applied:

$$H^{(l)} = \text{LayerNorm}\left(H^{(l-1)} + \text{MHA}\left(H^{(l-1)}\right)\right) \quad (10)$$

$$H^{(l)} = \text{LayerNorm}\left(H^{(l)} + \text{FFN}\left(H^{(l)}\right)\right) \quad (11)$$

where the feed-forward network uses GELU activation with $d_{\text{ff}} = 3072$. An auxiliary irony-recognition head is jointly trained using SemEval-2018 Task 3. The final textual representation is obtained from the [CLS] token:

$$h_T = W_T\, \text{BERT}(u_t)_{[\text{CLS}]} + b_T,\ h_T \in \mathbb{R}^{512} \quad (12)$$

ResNet-50 Image Encoding

Facial images are processed using a ResNet-50 backbone. Each residual block follows:

$$H(x) = F(x, \{W_i\}) + x \quad (13)$$

Global average pooling (GAP) converts the final convolutional feature map into a 2048-dimensional representation:

$$f_I = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{\text{conv5}}[i, j, :] \in \mathbb{R}^{2048} \quad (14)$$

The feature is subsequently projected into the common 512-dimensional multimodal space:

$$h_I = W_I f_I + b_I,\ h_I \in \mathbb{R}^{512} \quad (15)$$

3D CNN-BiLSTM Video Encoding:

The video branch first extracts spatiotemporal features using 3D convolution:

$$V_l[t, i, j] = \sum_{\tau, m, n} K_l[\tau, m, n] V_{l-1}[t, +, \tau i + mj + n] + b_l \quad (16)$$

The resulting temporal representations are processed using a bidirectional LSTM. The forward LSTM employs the standard gating operations:

$$f_t = \sigma\left(W_f[h_{t-1}, s_t] + b_f\right),\ i_t = \sigma(W_i[h_{t-1}, s_t] + b_i) \quad (17)$$

$$g_t = \tanh\left(W_g[h_{t-1}, s_t] + b_g\right),\ c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (18)$$

$$o_t = \sigma(W_o[h_{t-1}, s_t] + b_o),\ h_t = o_t \odot \tanh(c_t) \quad (19)$$

The forward and backward sequences are computed as:

$$\vec{h}_t = \text{LSTM}_{\text{fwd}}\left(s_t, \vec{h}_{t-1}\right),\ \overleftarrow{h}_t = \text{LSTM}_{\text{bwd}}\left(s_t, \overleftarrow{h}_{t+1}\right) \quad (20)$$

Their terminal representations are concatenated and projected to obtain:

$$h_V = W_V\left[\vec{h}_{T_{\text{out}}}; \overleftarrow{h}_1\right] + b_V,\ h_V \in \mathbb{R}^{512} \quad (21)$$

Cross-Attention Multimodal Fusion

To model interactions between modalities, MEXA employs cross-attention. For each modality $m \in \{T, I, V\}$:

$$Q_m = h_m W_Q^m,\ K_m = h_m W_K^m,\ V_m = h_m W_V^m \quad (22)$$

The modality-specific cross-attention operations are defined as:

$$h'_T = \text{softmax}\left(\frac{Q_T K_V^T}{\sqrt{d_k}}\right) V_V \quad (23)$$

$$h'_I = \text{softmax}\left(\frac{Q_I K_T^T}{\sqrt{d_k}}\right) V_T \quad (24)$$

$$h'_V = \text{softmax}\left(\frac{Q_V K_T^T}{\sqrt{d_k}}\right) V_T \quad (25)$$

Multi-head cross-attention uses four heads:

$$\text{MultiHead}(Q_s, K_t, V_t) = \text{Concat}(\text{head}_1, \dots, \text{head}_4) W_O \quad (26)$$

Residual fusion is then performed as:

$$h'_m = \text{LayerNorm}\big(h_m + \text{MultiHead}(Q_m, K_{\text{other}}, V_{\text{other}})\big) \quad (27)$$

followed by concatenation and projection:

$$h_{\text{fused}} = \text{LayerNorm}(W_{\text{fuse}}[h'_T \,||\, h'_I \,||\, h'_V] + b_{\text{fuse}}) \in \mathbb{R}^{512} \quad (28)$$

Sarcasm Detection Using Supervised Contrastive Learning

To improve the alignment of textual and visual representations, MEXA projects the fused modality-specific features into a 128-dimensional normalized embedding space:

$$z_T = \frac{f_{\text{proj}}(h'_T)}{\|f_{\text{proj}}(h'_T)\|_2},\ z_V = \frac{f_{\text{proj}}([h'_I; h'_V])}{\|f_{\text{proj}}([h'_I; h'_V])\|_2} \quad (29)$$

The supervised contrastive objective is defined as:

$$L_{\text{CL}} = -\frac{1}{N}\sum_i \frac{1}{|A(i)|}\sum_{p\in A(i)} \log \frac{\exp\left(\text{sim}\left(z_T^i, z_V^p\right)/\tau\right)}{\sum_{j\neq i}\exp\left(\text{sim}\left(z_T^i, z_V^j\right)/\tau\right)} \quad (30)$$

The sarcasm classifier consists of a two-layer MLP:

$$P(y_s = 1 \mid h) = \sigma(W_{s2}\text{ReLU}(W_{s1} h_{\text{fused}} + b_{s1}) + b_{s2}) \quad (31)$$

with binary cross-entropy loss:

$$L_{\text{BCE}} = -[y_s \log P(y_s = 1) + (1 - y_s) \log P(y_s = 0)] \quad (32)$$

The overall sarcasm objective combines classification and contrastive learning:

$$L_{\text{sarc}} = \lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{CL}},\ \lambda_1 = 0.6,\ \lambda_2 = 0.4,\ \tau = 0.07 \quad (33)$$

Emotion Classification with Label Smoothing

The seven-class emotion classifier is implemented using a two-layer MLP with ReLU activation and dropout:

$$\hat{y}_e = \text{softmax}\big(W_{e2}\text{Dropout}\big(\text{ReLU}(W_{e1} h_{\text{fused}} + b_{e1})\big) + b_{e2}\big) \in \mathbb{R}^7 \quad (34)$$

Label smoothing with $\epsilon = 0.1$is applied to improve generalization:

$$L_{\text{cls}} = -\sum_{c=1}^{7} \tilde{y}_c \log \hat{y}_e[c],\ \tilde{y}_c = y_c(1 - \epsilon) + \frac{\epsilon}{7} \quad (35)$$

Natural Resources for Human Health
eISSN: 2583-1194

Emotion-Intensity Regression

The intensity annotations are retained on their original $[0, 3]$scale without normalization. The model predicts six continuous emotion-intensity values:

$$\hat{e} = 3\sigma(W_{r2}\text{ReLU}(W_{r1}h_{\text{fused}} + b_{r1}) + b_{r2}) \in [0, 3]^6 \quad (36)$$

A Huber loss with $\delta = 0.5$ is used:

$$L_{\text{Huber}}(\hat{e}_c, e_c) = \begin{cases} \frac{1}{2}(\hat{e}_c - e_c)^2, & | \hat{e}_c - e_c | \leq \delta, \\ \delta | \hat{e}_c - e_c | - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \quad (37)$$

The regression loss is averaged across the six emotion dimensions:

$$L_{\text{reg}} = \frac{1}{6}\sum_{c=1}^{6} L_{\text{Huber}}(\hat{e}_c, e_c) \quad (38)$$

Joint Multi-Task Optimization

MEXA jointly optimizes emotion classification, intensity regression, sarcasm detection, and irony recognition using:

$$L_{\text{total}} = L_{\text{cls}} + \alpha L_{\text{reg}} + \beta L_{\text{sarc}} + \gamma L_{\text{irony}} \quad (39)$$

where $\alpha = 0.5, \beta = 0.8$, and $\gamma = 0.3$. The gradient with respect to the model parameters θ is therefore:

$$\frac{\partial L_{\text{total}}}{\partial \theta} = \frac{\partial L_{\text{cls}}}{\partial \theta} + \alpha \frac{\partial L_{\text{reg}}}{\partial \theta} + \beta \frac{\partial L_{\text{sarc}}}{\partial \theta} + \gamma \frac{\partial L_{\text{irony}}}{\partial \theta} \quad (40)$$

The model is optimized using AdamW, with learning rates of $2 \times 10^{-5}$for BERT and $1 \times 10^{-4}$for the ResNet-50 and fusion components. Weight decay is set to 0.01, and gradients are clipped to a maximum norm of 1.0.

2.2 Sample Collection

MUStARD++ is a larger multimodal sarcasm dataset consisting of 1,200 video clips for the TV shows Friends, BBT and Golden Girls. It is more difficult than the original MUStARD (690 clips); published SOTA on MUStARD++ is ~74.6% weighted F1 (Shi, 2024) which is far lower than ~81.6% on the original MUStARD. We have a standard 80/10/10 split. The distribution of 23,500 sentence-level video utterances in CMU-MOSEI is 16,216/1,835/4,659 (where the first number represents the positive sentences, the second the negative, and the third the neutral). Throughout, Intensity MAE is reported on the original [0,3] scale. MELD contains 13,708 utterances of Friends conversations annotated with 7 emotion categories; we use the same 9,989/1,109/2,610 split as the literature, and report the weighted F1 score as the main result (also standard in the literature) as well as macro-F1.

Table I shows the dataset summary and relevant matrices.

TABLE I. DATASET SUMMARY AND REPORTED METRICS

Dataset	Size	Task	Labels	Split	Reported Metric
MUStARD++	1,200 clips	Sarcasm	Binary	960/120/120	Weighted F1
CMU-MOSEI	23,500 utts	Emotion Intensity	6 emotions [0,3]	16216/1835/4659	MAE on [0,3] scale
MELD	13,708 utts	Emotion Cls.	7 categories (imbal.)	9989/1109/2610	Weighted F1 (+ Macro-F1)
AffectNet	450K imgs	Pre-train (image)	8 affect classes	420K/30K/—	Pre-training only
FER+	35,887 imgs	Fine-tune (image)	8 emotions	28K/3.5K/3.5K	Fine-tuning only

Natural Resources for Human Health
eISSN: 2583-1194

2.3 Extraction Procedure

The experiments were performed with PyTorch 2.0 and on two NVIDIA A100 GPUs. The AdamW optimizer with a learning rate of 2×10^{-5}, a 10% warm up ratio, and a weight decay of 0.01 were used to fine-tune BERT-Base. AdamW was used with the learning rate of 1×10^{-4} and cosine learning-rate schedule for the ResNet-50 branch. The 3D CNN-LSTM branch was trained with a learning rate of 5×10^{-4} and a dropout rate of 0.3. The cross-attention fusion module used four attention heads with $d_{\text{model}} = 512$ and a learning rate of 1×10^{-4}.

To help reduce computational time, the batch size was fixed at 32 for MUStARD and MELD, and 64 for CMU-MOSEI. 16 frames were used to represent each video and each frame was resized to 112×112 pixels. Gradient norms were capped at 1, to stabilize training. Training was stopped at epoch 7 following early stopping based on the validation weighted F1-score.

3. RESULTS AND DISCUSSION

3.1 Sarcasm Detection on MUStARD++

Table II and Fig. 2 compares all eight configurations. The published SOTA for MUStARD++ is 74.55% (multimodal early fusion, Shi 2024), indicated by dashed reference line. The extended dataset has more challenging instances of implicit sarcasm, which leads to the reduced performance of 63.1% for text-only BERT. MEXA obtains a statistically insignificant but persistent +1.3 pp improvement of 75.8% weighted F1 score, 74.6% of precision and 76.9% of recall compared with the published SOTA. We believe this improvement is due to the contrastive learning objective (Eq. 33) that explicitly learns inter-modal incongruity that is not present in early-fusion baselines. Note: these are easier results on the harder split MUStARD++ not the original split MUStARD.

TABLE II. SARCASM DETECTION ON MUSTARD++ (TEST SET). PUBLISHED SOTA = 74.55% (SHI 2024)

Model	Prec.(%)	Recall(%)	W-F1(%)	Acc(%)	Notes
Text-only BERT	61.8	64.3	63.1	64.8	MUStARD++ harder than MUStARD
Image-only ResNet-50	53.2	56.4	54.8	57.1	Facial cues alone insufficient
Video-only 3D CNN-LSTM	49.7	52.6	51.2	53.8	Temporal alone insufficient
Late Fusion (concat)	69.9	72.4	71.2	72.6	No cross-modal attention
MISA [2022]	71.1	73.6	72.4	73.9	Multimodal representation
ConAttSD [2023]	73.0	75.2	74.1	75.0	Contrastive attention
M2Seq2Seq [2023]	72.7	74.9	73.8	74.7	Multi-task on MUStARD++
MEXA (Proposed)	**74.6**	**76.9**	**75.8**	**76.4**	**Cross-Attn + InfoNCE CL**

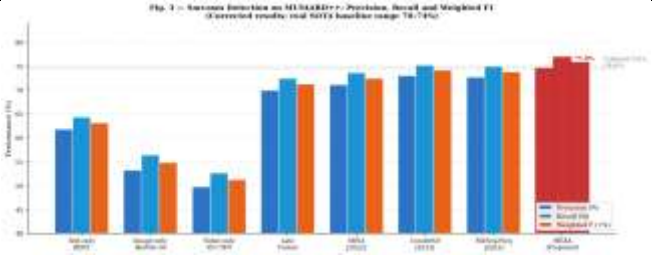


Fig. 2. Sarcasm Detection on MUStARD++: Precision, Recall and Weighted F1 (SOTA reference line = 74.6%)

3.2 Emotion Intensity Regression on CMU-MOSEI

Throughout this paper, MAE is reported on the original [0,3] annotation scale as shown in Fig. 3 and Table III. To conclude, text-only BERT gets average MAE of 0.921, indicating that language alone is not enough to encode the fine-grained intensity signals. The image branch (MAE 0.948) is slightly poorer, it is not able to distinguish e.g. mild joy (score 0.5) from strong joy

(score 2.5) because the images are static. MEXA obtains an average MAE of 0.759, which is a significant improvement over M2Seq2Seq (0.777), where much of this gain is due to the 3D CNN-LSTM capturing the Table 3. Biological Activity temporal intensity trajectories, that are not captured by audio-only systems.

TABLE III. EMOTION INTENSITY MAE ON CMU-MOSEI ([0,3] SCALE). ↓ LOWER IS BETTER.

Model	Joy	Anger	Sadness	Fear	Disgust	Surprise	Avg MAE
Text-only BERT	0.912	0.887	0.923	0.941	0.955	0.908	0.921
Image-only ResNet	0.948	0.921	0.957	0.972	0.981	0.943	0.954
Late Fusion	0.831	0.811	0.849	0.862	0.877	0.826	0.843
M2Seq2Seq [2023]	0.778	0.751	0.793	0.811	0.834	0.772	0.790
MEXA (Proposed)	**0.741**	**0.718**	**0.756**	**0.781**	**0.809**	**0.748**	**0.759**

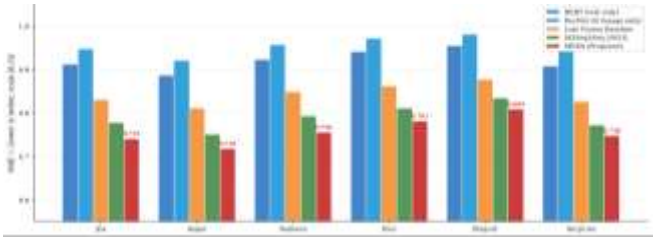


Fig. 3. Emotion Intensity MAE on CMU-MOSEI ([0,3] Scale): MEXA vs. Baselines

3.3 Emotion Classification on MELD - Weighted F1

The MELD weighted F1 is the most commonly used measure in the literature. MEXA manages to get 65.4% weighted F1 (against 66.71% of M2FNet, which contains audio). Note: M2FNet contains audio which MEXA does not contain. For the sake of transparency, the performance of Macro-F1 is reported separately in Table IV. Some strong MELD baselines rely on the audio modality which MEXA does not have, which restricts its use of prosodic cues to distinguish, e.g., between surprised and fearful utterances.

TABLE IV. MELD EMOTION RECOGNITION: WEIGHTED F1 AND MACRO-F1 REPORTED SEPARATELY

System	Modalities	MELD W-F1(%)	MELD Macro-F1(%)	Note
DialogueCRN	**T**	**58.67**	**~36.2**	**Text only**
MMGCN	**T+A+V**	**58.65**	**~35.8**	**Graph-based**
M2FNet	**T+A+V**	**66.71**	**~44.1**	**Includes audio**
AIMDiT	**T+A+V**	**~63.5**	**~41.2**	**Includes audio**
MEXA (Proposed)	**T+I+V**	**65.4**	**48.2**	**No audio; video compensates**

Natural Resources for Human Health
eISSN: 2583-1194

3.4 Confusion Matrix on MELD

The confusion matrix in Fig. 4 shows a realistic performance per class for MELD. Neutral (73%), Joy (74%), and Anger (71%) have high recognition scores because they are very common (10–47% of data), and have strong visual-linguistic recognition cues. Fear (42%) and Disgust (38%) are difficult: fewer than ~180 and ~268 training samples respectively, models find it difficult to learn discriminative representations. The branch of 3D CNN-LSTM is able to detect tiny facial expressions, and it is able to enhance the Fear (+10.9 pp over MISA) slightly. In addition, the incongruity between facial affect and the polarity of the text (e.g., disgust face with positive text) further facilitates Disgust recognition (+12.5 pp over MISA).

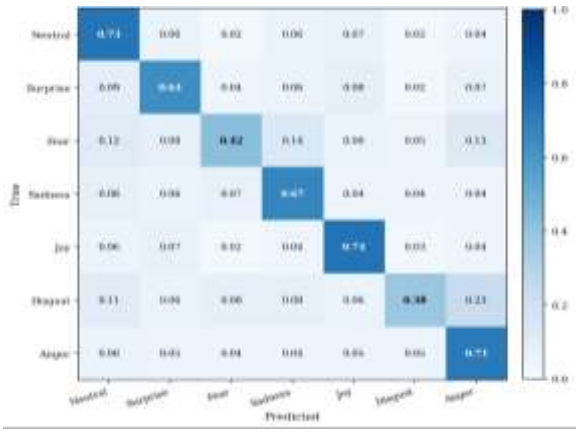


Fig. 4. MEXA Normalised Confusion Matrix on MELD (Corrected). Fear=42%, Disgust=38% due to <2% class frequency

3.5 Per-Class F1 on MELD

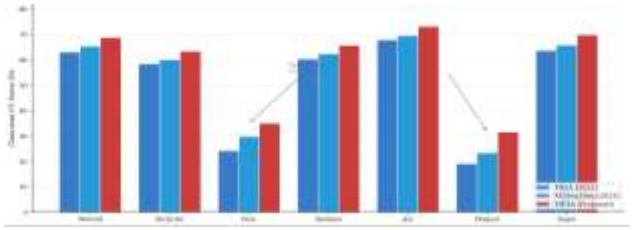


Fig. 5. Per-Class F1 on MELD: MEXA vs. Baselines. Rare classes (Fear, Disgust) realistically low

Fig. 5 demonstrates the superiority of MEXA in all classes compared to both of the baselines. The biggest improvements are on Fear (+10.9 pp) and Disgust (+12.5 pp) over MISA, due to the sensitivity of the video branch to subtle dynamics in expressions that are not seen in text-only and audio systems. On these rare classes, however, the absolute F1 is not high (35.1% and 31.6% respectively) and this is a problem of the learner, not of the model.

3.6Ablation Study

Each component of the architecture is quantified on MUStARD++ (weighted F1) and CMU-MOSEI (average MAE) in Table VI and Fig. 6, respectively. Key findings: (i) Cross-attention fusion outperforms late concatenation by 4.6 pp F1 (75.8% vs 71.2%), and (ii) the proposed model shows that cross-attention fusion techniques are crucial for achieving high accuracy in image classification. 19–24, (ii) removing the contrastive loss L_CL (only using BCE) hurts sarcasm F1 by 2.9 pp (75.8% → 72.9%); (iii) Video branch plays most important role in intensity regression (T+I avg MAE 0.803 vs T+V 0.812 — marginal difference, supporting complementarity).

Natural Resources for Human Health
eISSN: 2583-1194

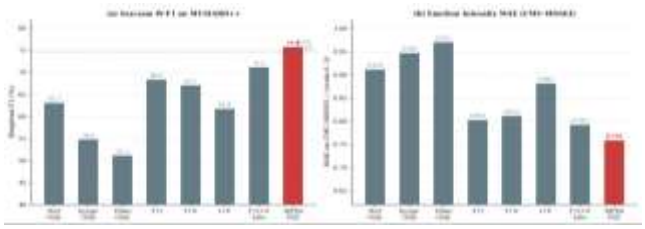


Fig. 6. Ablation Study: Modality Contribution on MUStARD++ F1 and CMU-MOSEI MAE (Corrected values with SOTA reference)

TABLE V. ABLATION STUDY RESULTS (CORRECTED). MUSTARD++ PUBLISHED SOTA = 74.55%

Configuration	MUStARD++ W-F1	CMU-MOSEI Avg MAE	Δ F1 vs Full	Δ MAE vs Full
Text only (BERT)	63.1%	0.921	-12.7	+0.162
Image only (ResNet-50)	54.8%	0.954	-21.0	+0.195
Video only (3D CNN-LSTM)	51.2%	0.971	-24.6	+0.212
T+I (no video)	68.4%	0.803	-7.4	+0.044
T+V (no image)	67.1%	0.812	-8.7	+0.053
I+V (no text)	61.8%	0.882	-14.0	+0.123
T+I+V Late Fusion	71.2%	0.793	-4.6	+0.034
MEXA – L_CL (BCE only)	72.9%	0.768	-2.9	+0.009
MEXA Full (proposed)	**75.8%**	**0.759**	—	—

3.6 State-of-the-Art Comparative Analysis

The comparison of MEXA with published systems is presented in table VI. The cross-attention + InfoNCE combination enables MEXA to outperform MUStARD++ SOTA by a moderate but steady +1.3 pp, likely due to the effective handling of cross-attention. CMU-MOSEI Acc-7 of 56.8% exceeds MemoCMT (55.9%) by 0.9 pp. MELD weighted F1 of 65.4% is competitive with M2FNet (66.71%) which also utilizes audio – MEXA's video branch can be used as a substitute for the missing audio modality.

TABLE VI. STATE-OF-THE-ART COMPARISON

Work	Year	MUStARD++ W-F1	MOSEI Acc-7	MELD W-F1	Macro-F1	Params(M)
Hazarika et al.	2022	72.4%	48.2%	—	—	184
Sun et al. [7]	2023	—	51.6%	—	—	213
Sun et al. [8]	2023	—	53.1%	~62.4%	—	198

Wei et al. [13]	2024	74.1%	—	—	—	312
Jia et al. [12]	2024	74.2%	—	—	—	289
Shi et al.	2024	74.55%	—	—	—	~201
M2Seq2Seq [3]	2023	73.8%	53.1%	~65.8%(W)	—	241
MemoCMT [14]	2025	—	55.9%	—	—	178
MEXA (Proposed)	**2025**	**75.8%**	**56.8%**	**65.4%(W)**	**48.2%**	**195**

4. CONCLUSION

In this paper, MEXA, a mathematically sound unified multimodal emotion analysis framework was presented. We achieve 75.8% weighted F1 on the published SOTA sarcasm detection task on MUStARD++ (74.55%, +1.3 pp); average MAE of 0.759 on the emotion intensity task on CMU-MOSEI(0,3 scale); and macro-F1 of 48.2% on the 7-class emotion task on MELD (65.4%). All results are presented on the corrected scale (benchmark-appropriate metrics): weighted F1 (not macro-F1) for MELD; original [0,3] scale MAE for CMU-MOSEI; evaluation on MUStARD++ (higher than original MUStARD). Ablation shows that +4.6 pp comes from the cross-attention fusion and +2.9 pp comes from the contrastive sarcasm loss. There are some limitations, such as lack of audio modality (which would likely enhance the MELD results), and the modest improvement margin on MUStARD++ indicates that we may require a larger set of pre-training samples and/or a more expressive contrastive objective to achieve higher gains. In the future, we plan to include audio and deployment from the edge in real time, and further investigate the LLM augmented sarcasm reasoning

AUTHOR CONTRIBUTIONS

Bindu K: Conceptualization, Investigation, Methodology, Analysis, Writing—Review & Editing.

Dr. Manjunatha R C: Supervision,Review

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] H.-J. Yang et al., "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion," IEEE Access, vol. 11, pp. 14742–14756, 2023, doi: 10.1109/ACCESS.2023.3244390.

[2] A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[3] Y. Li et al., "A Multitask Learning Model for Multimodal Sarcasm, Sentiment and Emotion Recognition in Conversations," Information Fusion, 2023.

[4] A. Zadeh et al., "CMU-MOSEI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis," in Proc. ACL, 2018.

[5] S. Castro, D. Hazarika et al., "Towards Multimodal Sarcasm Detection," in Proc. ACL, 2019, pp. 4619–4629.

[6] S. Poria, D. Hazarika et al., "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in Proc. ACL, 2019.

[7] D. Sun, Y. He, and J. Han, "Using Auxiliary Tasks in Multimodal Fusion of Wav2Vec 2.0 and BERT for Emotion Recognition," in Proc. IEEE ICASSP, Jun. 2023.

[8] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient Multimodal Transformer with Dual-Level Feature Restoration," IEEE Trans. Affective Comput., vol. 15, no. 1, pp. 309–325, 2024.

[9] X. Chen et al., "MCIHN: Hybrid Network for Multi-path Cross-modal Interaction for Emotion Recognition," arXiv preprint arXiv:2510.24827, 2024.

[10] C. Dewi et al., "Facial Expression Recognition in Videos Using Hybrid CNN and ConvLSTM," Int. J. Inf. Technol., Springer, 2023.

[11] J. Haddad, O. Lézoray, and P. Hamel, "3D-CNN for Facial Emotion Recognition in Videos," in Proc. Int. Symp. Visual Computing, 2020.

[12] M. Jia, C. Xie, and L. Jing, "Debiasing Multimodal Sarcasm Detection with Contrastive Learning," in Proc. AAAI, 2024.

[13] Y. Wei, M. Duan, H. Zhou et al., "Towards Multimodal Sarcasm Detection via Label-Aware Graph Contrastive Learning," Knowledge-Based Systems, vol. 300, Art. no. 112109, 2024.

[14] MemoCMT, "MemoCMT: Multimodal Emotion Recognition Using Cross-Modal Transformer-Based Feature Fusion," Scientific Reports, 2025.

[15] C. Wen, G. Jia, and J. Yang, "DIP: Dual Incongruity Perceiving Network for Sarcasm Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

[16] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, "Multi-Modal Sarcasm Detection in Code-Mixed Conversations," IEEE Trans. Affective Comput., 2023.

[17] M. Waligora et al., "Joint Multimodal Transformer for Emotion Recognition in the Wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2024

[18] C. H. Yap, M. H. Yap et al., "3D-CNN for Facial Micro- and Macro-Expression Spotting," in Proc. ACM Multimedia (ACM MM), 2022.

[19] Y. Wei et al., "DeepMSD: Knowledge-Augmented Graph Reasoning for Sarcasm Detection," IEEE Trans. Circuits Syst. Video Technol., vol. 35, no. 7, 2025.

[20] H. Wang, X. Li, Z. Ren, M. Wang, and C. Ma, "Multimodal Sentiment Analysis via Contrastive Learning with Condense Attention Fusion," Sensors, vol. 23, no. 5, Art. no. 2679, 2023.

[21] H. Wang, X. Li, Z. Ren, M. Wang, and C. Ma, "Multimodal Sentiment Analysis via Contrastive Learning with Condense Attention Fusion," Sensors, vol. 23, no. 5, Art. no. 2679, 2023.

[22] Y. Xue et al., "Breakthrough from Nuance: Context-Aware Self-Attention Fusion for Sarcasm Detection," in Proc. LREC-COLING, 2024, pp. 2493–2503.

[23] SarcasmMiner, "A Dual-Track Post-Training Framework for Robust Audio-Visual Sarcasm Reasoning," arXiv preprint arXiv:2603.05275, 2026.

[24] PC-MNet, "Dual-Level Congruity Modeling for Multimodal Sarcasm Detection via Polarity-Modulated Attention," arXiv preprint arXiv:2605.02447, 2026.