

A Bidirectional Cross-Modal Attention Framework for Multimodal Sentiment Analysis

B S Harish, C K Roopa, M S Kendagannaswamy, T S Kaveri*

Department of Information Science and Engineering, JSS Science and Technology University, Mysuru 570006, Karnataka, India

Abstract—Multimodal sentiment analysis on social media data presents unique challenges due to label noise, modality conflicts, and class imbalance inherent in annotated image-text datasets. The proposed work presents CLIP-CrossFusion Net, a novel multimodal sentiment analysis framework that integrates the Contrastive Language-Image Pre-Training (CLIP) ViT-B/32 image encoder with a RoBERTa-Base text encoder through a bidirectional cross-modal attention mechanism. The proposed approach addresses three key challenges: 1) rich cross-modal feature interaction via bidirectional attention, 2) noisy label robustness through focal loss with label smoothing, and 3) class imbalance mitigation via targeted class weighting and Mixup augmentation. The study further proposes an extended variant with a colour feature branch, supervised contrastive loss, and deeper cross-modal attention layers. Using the proposed model, comprehensive experiments are conducted on both the MVSA-Multiple (19,600 image-text tweet pairs) and MVSA-Single (4,511 pairs) datasets. On MVSA-Multiple, the proposed model achieves 74.39% accuracy and a weighted F1 of 0.7497. On MVSA-Single, the proposed ensemble of three complementary model variants (base model, colour-contrastive variant, and RoBERTa-Large variant) achieved 75.66% accuracy and an F1 of 0.7517, outperforming strong baselines. These results demonstrate the effectiveness of cross-modal attention fusion combined with noise-robust training strategies for multimodal sentiment analysis.

Keywords—Multimodal sentiment analysis; CLIP; RoBERTa; cross-modal attention; social media; deep learning; transformer

I. INTRODUCTION

Twitter, Instagram, and Weibo generate colossal quantities of multimodal content, posts in which opinions are conveyed in the form of text, images, and video simultaneously. Sentiment analysis of such content can be helpful in brand monitoring, tracking health and politics, as well as financial projections [1]. Text-only strategies are very good; however, they still lack what images are saying and the interaction between text and images.

Combining text and image processing (MSA) is more effective than either of the two, yet there are real challenges. Social media image-text pairs often contain conflicting sentiment, e.g., a tweet may contain a complaint, but the image will appear cheerful. To further complicate matters, the MVSA-Multiple dataset [2] has a lot of label noise, because of disagreements between its three annotators. The most difficult of all cases is the Neutral category, as the neutral sentiment is context-specific and can be misinterpreted as positive or negative.

As an extension of this work, this study introduces CLIP-CrossFusion Net, a new multimodal sentiment analysis model, which has three main contributions:

- A cross-modal attention process in both directions, with text and image features cross-attending in parallel so that each modality can influence the other to be reshaped and then fused.
- A noise-robust training method based on Focal Loss and class-balanced weights, label smoothing, and Mixup augmentation to deal with label noise in MVSA-Multiple and improve the performance of Neutral classes.
- Ablation experiments on variants of models in simple concatenation baselines and full cross-modal transformer architectures, and the way each component contributes to the final output.

The proposed MVSA-Multiple experiments achieve an accuracy of 74.39% and a weighted F1 of 0.7497, which is higher than strong baselines. The Neutral class F1, which is enhanced by the weighting of classes, improves from 0.59 to 0.63.

II. RELATED WORK

A. Unimodal Sentiment Analysis

Sentiment analysis has a rich history dating back to lexicon-based approaches [3]. CNN-based [4] and LSTM-based [5] architectures improved performance, followed by attention mechanisms [6] and transformer-based models such as BERT [7] and RoBERTa [8]. However, unimodal approaches are fundamentally limited for social media content where images carry substantial sentiment information.

B. Multimodal Sentiment Analysis

Early multimodal approaches were based on the concatenation of simple features of CNN images and word embeddings [9]. Graph neural network models like MGNNS represented cross-modal relationships as graphs [10]. The D&R Net [11] was able to determine the cross-modal contrast as well as semantic correlation by decomposition and relational networks. CLMLF [12] used transformer-fusion CNN-supervised contrastive learning. MVSA-Single with cross-modal interaction fusion on SECIF [13] produced good results. ICCI [14] scored 79.33% on MVSA-Single by combining HSV colour cues with semantic features.

*Corresponding author.

Using VGG16, Mask-RCNN, BERT, WCNNE, and Dempster-Shafer theory, Biswas et al. [15] obtained 93.48% on MVSA-Multiple.

C. Vision-Language Pretraining

CLIP [16] proposed contrastive pretraining using 400M image-text pairs, and it trains a shared embedding space with impressive transfer properties. ViT [17] showed image recognition with pure transformer architectures, and DeiT [18] optimized vision transformers with distillation. Such models offer much more detailed representations than the conventional CNN encoders, which is why they are adopted into the proposed framework.

D. Cross-Modal Attention and Colour Features

Deep cross-modal attention was applied to two-stage feature fusion in DTN [19]. The SRC-Model [13] used bidirectional image-text interaction gating and MHA. MIGSIE [13] applied GNN-based text-guided multi-channel modules. ICCI showed that the colour cues presented by images are good sentiment indicators, which drives the colour feature branch. The main task of this work is to combine CLIP joint pretraining with explicit cross-modal attention and colour features.

III. DATASETS

A. MVSA-Multiple Dataset

The MVSA-Multiple dataset [2] comprises 19,600 image-text pairs on Twitter that have been annotated by three human raters. The final joint label is obtained by majority voting, and the conflicting pairs are eliminated. It has a large class imbalance: Positive (42.0%), Negative (31.4%), Neutral (26.7%), and inherent label noise due to annotation disagreement. Fig. 1 shows the distribution of the classes in the train, validation, and test splits, with the positive class being predominant and the relatively low representation of the neutral category, which compels the class-weighting approach.

B. MVSA-Single Dataset

The MVSA-Single dataset is a set of 4,511 valid image-text pairs on Twitter (after filtering) annotated by one rater. The extreme imbalance of classes, Positive (59.5%), Negative (30.1%), Neutral (10.4%), is problematic with the neutral category. Hence, a typical 8:1:1 train/validation/test split with stratified sampling was applied. This extreme imbalance is depicted visually in Fig. 2, where positive samples are nearly 60% of the data and neutral samples are only 10.4%. Thus, correct Neutral classification is the most difficult on this dataset. Table I summarises the statistics of both datasets.

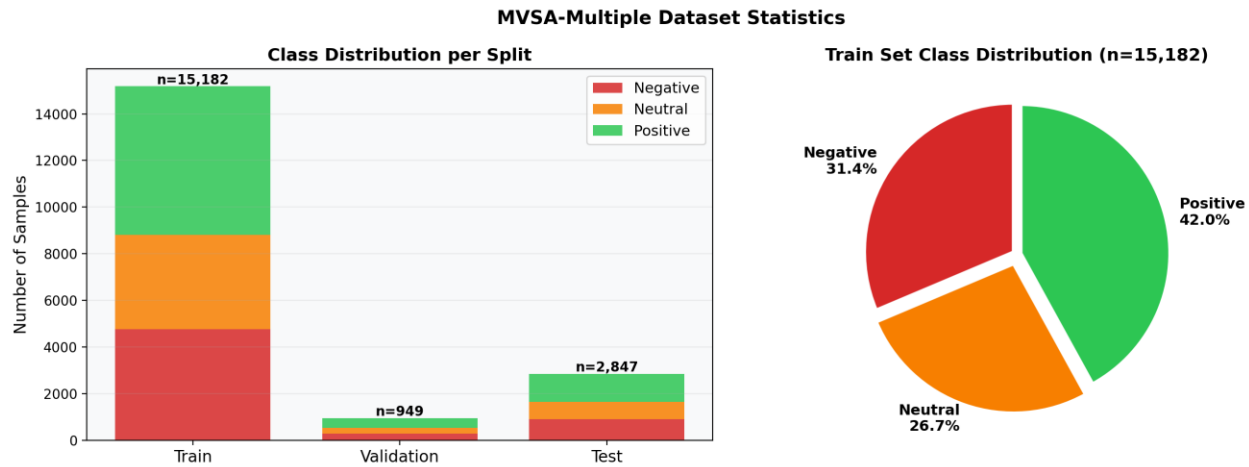


Fig. 1. MVSA-Multiple dataset class distribution across train, validation, and test splits.

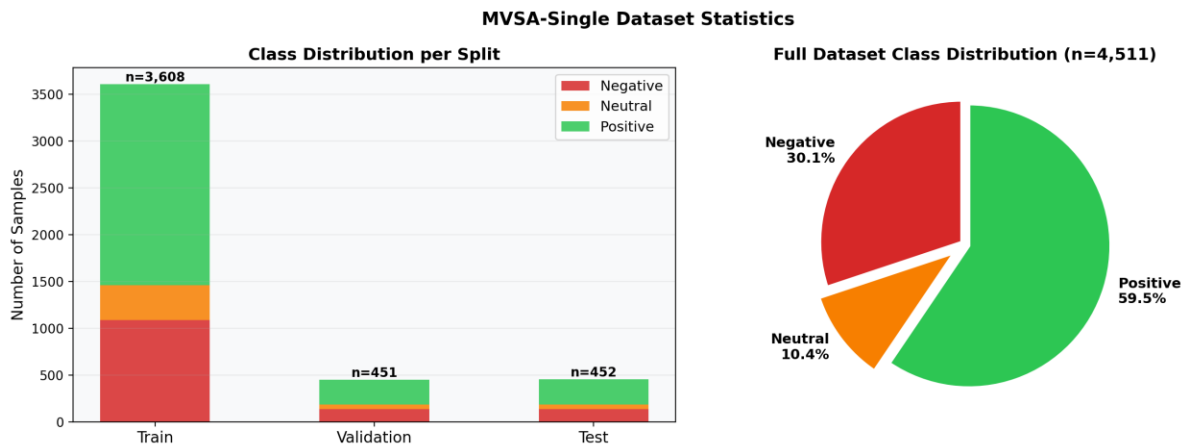


Fig. 2. MVSA-Single dataset class distribution. Severe positive dominance (59.5%) makes neutral classification challenging.

TABLE I. DATASET STATISTICS FOR MVSA-MULTIPLE AND MVSA-SINGLE (ANNOT. = NUMBER OF ANNOTATORS PER SAMPLE)

Dataset	Split	Neg	Neu	Pos	Total	Annot.	Split Ratio
MVSA-Multiple	Train	4,760	4,047	6,375	15,182	3	Pre-split
	Val	290	244	415	949		
	Test	893	759	1,195	2,847		
MVSA-Single	Train	1,086	376	2,146	3,608	1	8:1:1
	Val	136	47	268	451		
	Test	136	47	269	452		

IV. METHODOLOGY

CLIP-CrossFusion Net follows a dual-stream encoder architecture where text and image modalities are independently encoded, projected to a shared embedding space, refined through bidirectional cross-modal attention, and fused via a gated mechanism before classification. The extended colour-contrastive variant adds a colour feature branch that extracts HSV colour histograms from images and integrates them through additional cross-modal attention layers. Fig. 3 presents a complete overview of the architecture, depicting the dual-stream encoders, shared projection layers, bidirectional cross-modal attention module, gated fusion mechanism, and final three-class classifier.

A. Encoders

1) *Image Encoder: CLIP ViT-B/32*: The proposed approach employs the Vision Transformer ViT-B/32 component of CLIP [16], pretrained on 400 million internet-collected image-text pairs. The encoder produces a 512-dimensional embedding. All layers are frozen except the final two to four transformer blocks, which are fine-tuned with a small learning rate to adapt visual representations to sentiment-relevant features, as expressed in (1).

$$\mathbf{f}_{\text{img}} = \text{CLIP}_{\text{ViT}}(\mathbf{x}_{\text{img}}) \in \mathbb{R}^{512} \quad (1)$$

2) *Text Encoder: RoBERTa-Base / RoBERTa-Large*: For text encoding, RoBERTa-Base [8] is employed for the base, cross-modal attention, focal-loss, and colour-contrastive variants, or RoBERTa-Large for the large-encoder variant. Tweet text is tokenised to a maximum of 128 tokens. The [CLS] token is extracted as the sentence-level representation, as given in (2). The same selective fine-tuning strategy applies: all layers are frozen except the last two to four blocks.

$$\mathbf{f}_{\text{text}} = \text{RoBERTa}([\text{CLS}], \mathbf{x}_{\text{text}}) \in \mathbb{R}^{768} \text{ or } \mathbb{R}^{1024} \quad (2)$$

B. Modality Projection

Both modalities are projected to a common 256-dimensional space using separate projection heads with LayerNorm and GELU activation, as defined in (3) and (4).

$$\mathbf{p}_{\text{img}} = \text{LayerNorm}(\text{GELU}(\mathbf{W}_{\text{img}} \cdot \mathbf{f}_{\text{img}})) \in \mathbb{R}^{256} \quad (3)$$

$$\mathbf{p}_{\text{text}} = \text{LayerNorm}(\text{GELU}(\mathbf{W}_{\text{text}} \cdot \mathbf{f}_{\text{text}})) \in \mathbb{R}^{256} \quad (4)$$

C. Bidirectional Cross-Modal Attention

The core innovation is the bidirectional cross-modal attention module enabling each modality to attend to the other

simultaneously, as formulated in (5) and (6) for the image-to-text and text-to-image directions, respectively. Two to three such layers are stacked for hierarchical cross-modal reasoning.

$$\mathbf{a}_{\text{img}} = \text{LayerNorm}(\mathbf{p}_{\text{img}} + \text{Dropout}(\text{MHA}(\mathbf{p}_{\text{img}}, \mathbf{p}_{\text{text}}, \mathbf{p}_{\text{text}}))) \quad (5)$$

$$\mathbf{a}_{\text{text}} = \text{LayerNorm}(\mathbf{p}_{\text{text}} + \text{Dropout}(\text{MHA}(\mathbf{p}_{\text{text}}, \mathbf{p}_{\text{img}}, \mathbf{p}_{\text{img}}))) \quad (6)$$

D. Colour Feature Branch

Inspired by ICCI [14], which demonstrated that colour information improves sentiment analysis, the extended variant introduces a dedicated colour branch. HSV colour histograms (32 bins each for the H, S, and V channels, giving a 96-dimensional feature) are extracted from each image and processed through a two-layer MLP, as given in (7) and (8). Bidirectional cross-modal attention between colour and image features, defined in (9) and (10), allows the model to leverage emotional colour cues.

$$\mathbf{c} = \text{concat}(\text{hist}_H, \text{hist}_S, \text{hist}_V) / \Sigma \in \mathbb{R}^{96} \quad (7)$$

$$\mathbf{p}_{\text{colour}} = \text{ColourBranch}(\mathbf{c}) \in \mathbb{R}^{256} \quad (8)$$

$$\mathbf{p}_{\text{img}} = \text{Attn}_{\text{col} \rightarrow \text{img}}(\mathbf{p}_{\text{img}}, \mathbf{p}_{\text{colour}}) \quad (9)$$

$$\mathbf{p}_{\text{colour}} = \text{Attn}_{\text{img} \rightarrow \text{col}}(\mathbf{p}_{\text{colour}}, \mathbf{p}_{\text{img}}) \quad (10)$$

E. Gated Fusion

The base, cross-modal attention, and focal-loss variants use dual-stream gated fusion, while the colour-contrastive and large-encoder variants use triple-stream fusion including colour features, as expressed in (11)-(13). The fused representation is 512-dimensional for the dual-stream variants and 768-dimensional for the colour-contrastive variants.

$$\mathbf{fused} = \text{concat}(\mathbf{a}_{\text{img}}, \mathbf{a}_{\text{text}}, \mathbf{p}_{\text{colour}}) \quad (11)$$

$$\text{gate} = \text{Sigmoid}(\mathbf{W}_{\text{gate}} \cdot \mathbf{fused}) \quad (12)$$

$$\mathbf{z} = \mathbf{fused} \odot \text{gate} \quad (13)$$

F. Training Strategy

1) *Focal loss with label smoothing*: Standard cross-entropy fails on MVSA due to label noise and class imbalance. Focal Loss with label smoothing and class-balanced weights targeting the hard Neutral class is used, as given in (14).

$$\mathbf{L}_{\text{FL}} = -\sum_i \mathbf{w}_i \cdot (1 - \mathbf{p}_i)^\gamma \cdot \mathbf{y}_i^{\text{smooth}} \cdot \log(\mathbf{p}_i) \quad (14)$$

Here $\gamma = 2.0$, label smoothing $\varepsilon = 0.1$, and class weights are [Neg: 1.2-1.5, Neu: 1.8-2.0, Pos: 1.0]. The Neutral class receives higher weight to improve recall. The focal term $(1 - \mathbf{p}_i)^\gamma$ down-weights easy samples.

2) *Supervised contrastive loss*: The colour-contrastive variant adds a supervised contrastive loss term that pulls same-class embeddings closer while pushing different-class embeddings apart, improving class separation especially for the ambiguous Neutral class. The contrastive term and the combined objective are given in (15) and (16).

$$L_{\text{contrast}} = -\text{mean}_{\text{anchors}} [\Sigma(\text{positive pairs}) / \Sigma(\text{all pairs})] \quad (15)$$

$$L_{\text{total}} = L_{\text{focal}} + \lambda \cdot L_{\text{contrast}}, \quad \lambda = 0.05 \quad (16)$$

3) *Mixup augmentation*: Mixup [20] is applied to the image modality with $\lambda \sim \text{Beta}(0.2, 0.2)$, creating mixed image-label pairs that prevent overly certain predictions on noisy labels.

4) *Optimisation*: AdamW [21] is used with layer-wise learning rates: encoders (3-5e-6) and fusion head (8e-5 to 1e-4). Training uses cosine annealing with 10% warmup, gradient clipping (max = 1.0), and gradient accumulation (steps = 2-4, effective batch = 64). Early stopping patience is 5-6 epochs on validation F1. The complete hyperparameter configuration of each model variant is listed in Table II.

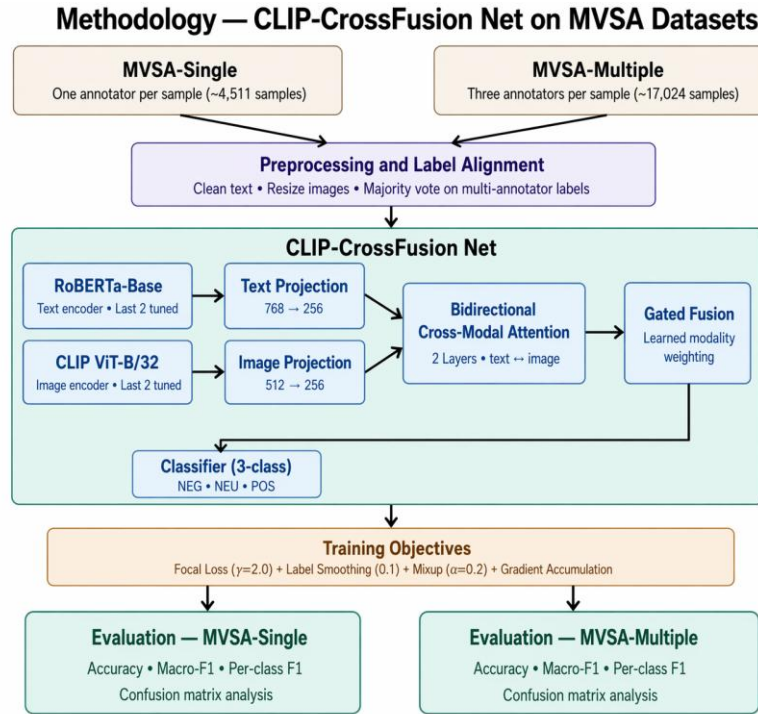


Fig. 3. CLIP-CrossFusion Net architecture: dual-stream encoders, projection, bidirectional cross-modal attention, gated fusion, and 3-class classifier.

TABLE II. HYPERPARAMETER CONFIGURATION PER MODEL VARIANT

Hyperparameter	Base / X-Attn / Focal	Colour + Contrastive	RoBERTa-Large Variant
Image encoder	CLIP ViT-B/32	CLIP ViT-B/32	CLIP ViT-B/32
Text encoder	RoBERTa-Base	RoBERTa-Base	RoBERTa-Large
Colour branch	No	Yes (96-dim)	Yes (96-dim)
Contrastive loss	No	Yes ($w = 0.05$)	Yes ($w = 0.05$)
Attention layers	2	3	3
Unfrozen layers	2	4	4
Focal loss γ	2.0	2.0	2.0
Class weights (Neg/Neu/Pos)	1.2 / 1.8 / 1.0	1.3 / 2.0 / 1.0	1.3 / 2.0 / 1.0
Hardware	T4 15 GB	T4 15 GB	T4 15 GB

G. Multi-Model Ensemble

For MVSA-Single, three complementary model variants are combined through weighted probability averaging, as given in (17). The base model provides strong text representations, the colour-contrastive variant adds colour information, and the RoBERTa-Large variant contributes larger text features. The

optimal weights (base: 0.5, colour-contrastive: 0.4, RoBERTa-Large: 0.1) are determined by grid search on the validation set. Test-Time Augmentation (three transforms: original, horizontal flip, centre crop) is applied for each model before ensembling.

$$P_{\text{ens}} = 0.5 \cdot P_{\text{base}} + 0.4 \cdot P_{\text{colour}} + 0.1 \cdot P_{\text{large}} \quad (17)$$

V. EXPERIMENTS

A. Experimental Setup

All experiments use PyTorch with mixed-precision training (FP16) on an NVIDIA Tesla T4 (15 GB VRAM). For the MVSA-Multiple dataset, the predefined train/validation/test splits are used. For MVSA-Single, 8:1:1 stratified random splits are used. Accuracy and weighted F1-score are reported as primary metrics, consistent with prior work.

B. Baselines

- DeiT-Base + BERT-Base (Concat): primary strong baseline using concatenation fusion.
- ViT-Base + BERT-Base (Concat): Vision Transformer with BERT.
- Swin-Base + BERT-Base (Concat): Swin Transformer with BERT.
- RoBERTa-Large + CLIP ViT-L/14 (large-encoder baseline): larger encoders with gated fusion.

- CLIP ViT-B/32 + RoBERTa-Base (cross-modal attention / focal-loss variants): cross-modal attention variants.
- ICCI (2023) [14]: cross-modal attention with colour cues; 79.33% on MVSA-Single and 73.29% on MVSA-Multiple.
- CLIP-CrossFusion with colour + contrastive (ours): with colour branch and contrastive loss.
- Ensemble of base + colour-contrastive + RoBERTa-Large (ours): best ensemble on MVSA-Single.

VI. RESULTS AND DISCUSSION

A. MVSA-Multiple Results

Table III presents results on MVSA-Multiple. The proposed CLIP-CrossFusion Net with Focal Loss and class weighting achieves 74.39% accuracy and 0.7497 weighted F1, outperforming all baselines including ICCI (73.29%). Fig. 4 visualises the comparison.

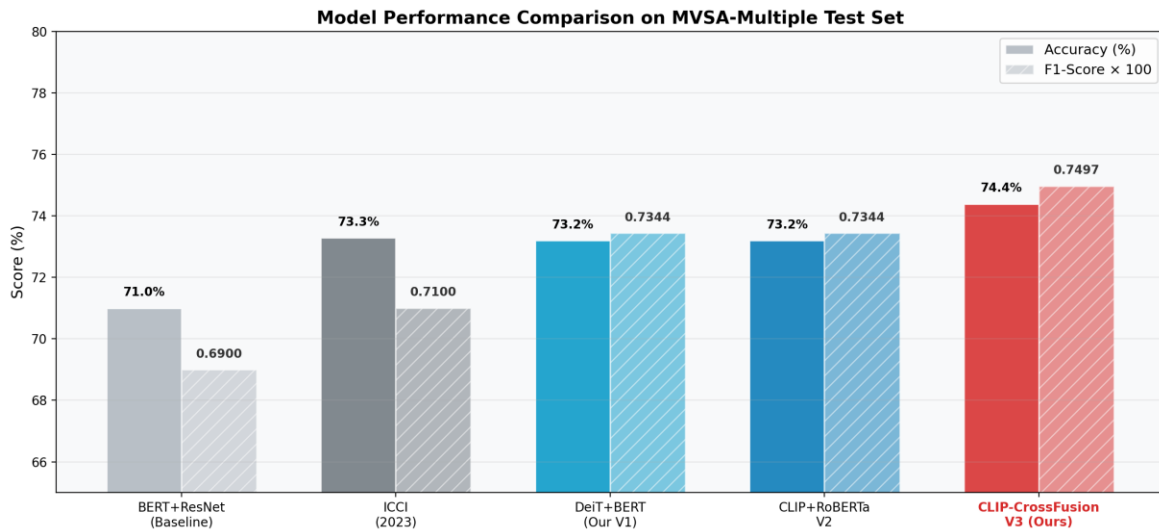


Fig. 4. Model performance comparison on the MVSA-Multiple test set. The proposed CLIP-CrossFusion Net with Focal Loss and class weighting achieves 74.39% accuracy.

TABLE III. PERFORMANCE COMPARISON ON THE MVSA-MULTIPLE TEST SET

Model	Acc. (%)	W-F1	Neg-F1	Neu-F1	Pos-F1
DeiT-Base + BERT (Concat)	58.76	0.5879	0.52	0.41	0.67
ViT-Base + BERT (Concat)	57.13	0.5669	0.49	0.39	0.65
Swin-Base + BERT (Concat)	56.63	0.5434	0.47	0.41	0.67
RoBERTa-L + CLIP ViT-L/14 (large-encoder)	71.13	0.7045	0.76	0.57	0.72
ICCI (2023) [14]	73.29	0.7100	–	–	–
CLIP ViT-B/32 + RoBERTa (cross-modal attention)	73.20	0.7344	0.84	0.59	0.75
CLIP-CrossFusion Net + Focal + CW (ours)	74.39	0.7497	0.85	0.63	0.75

B. MVSA-Single Results

Table IV presents results on MVSA-Single. The proposed ensemble of the base model, colour-contrastive variant, and RoBERTa-Large variant achieves 75.66% accuracy and 0.7517 weighted F1, outperforming DeiT+BERT (61.50%) and

substantially improving over the single best model (74.12%). The Neutral F1 improves from 0.38 (base model) to 0.45 (ensemble), a significant gain on the hardest class. Fig. 5 visualises this comparison, showing the progressive accuracy gains from the base model to the final ensemble, with the best result (75.66%) approaching the ICCI benchmark of 79.33%.

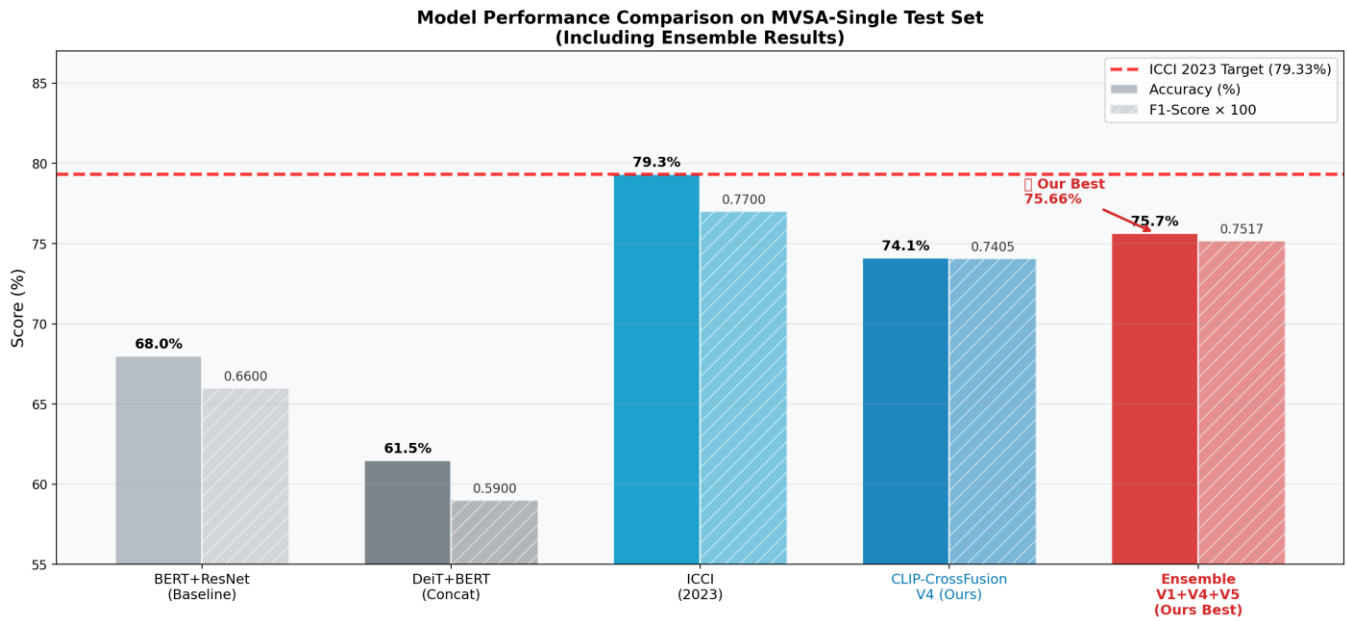


Fig. 5. Model performance comparison on the MVSA-Single test set.

TABLE IV. PERFORMANCE COMPARISON ON THE MVSA-SINGLE TEST SET

Model	Acc. (%)	W-F1	Neg-F1	Neu-F1	Pos-F1
BERT + ResNet (baseline)	68.00	0.6600	–	–	–
DeiT + BERT (Concat)	61.50	0.5900	–	–	–
ICCI (2023) [14]	79.33	0.7700	–	–	–
Base: CLIP + RoBERTa-Base	73.45	0.7288	0.65	0.38	0.83
+ Colour + Contrastive	74.12	0.7405	0.67	0.41	0.84
+ RoBERTa-Large	73.67	0.7355	0.67	0.40	0.83
Ensemble: base + colour-contrastive + large (ours)	75.66	0.7517	0.67	0.45	0.85

C. Ablation Study

Table V traces the contribution of each component across both datasets, showing systematic improvement as components are added.

TABLE V. ABLATION STUDY ACROSS BOTH DATASETS (X-ATTN = CROSS-MODAL ATTENTION, CW = CLASS WEIGHTS, CONT. = CONTRASTIVE LOSS)

Model	CLIP	X-Attn	Focal	CW	Colour	Cont.	Mult. (%)	Sing. (%)
DeiT + BERT	X	X	X	X	X	X	58.76	61.50
Base (CLIP + X-Attn)	✓	✓	X	X	X	X	71.13	73.45
+ Focal + CW	✓	✓	✓	✓	X	X	74.39	–
+ Colour + Contrastive	✓	✓	✓	✓	✓	✓	–	74.12
Ensemble (base + colour-contrastive + large)	✓	✓	✓	✓	✓	✓	–	75.66

D. Training Dynamics on MVSA-Multiple

Fig. 6 and Fig. 7 illustrate the training dynamics on MVSA-Multiple. Fig. 6 shows the overall accuracy and loss convergence, with the best checkpoint reached at epoch 10

(validation accuracy 74.71%, validation F1 0.7503). Fig. 7 tracks per-class F1 progression and demonstrates how targeted class weighting steadily improves the Neutral class F1 from 0.55 to 0.63 over training.

CLIP-CrossFusion Net — Training Curves (MVSA-Multiple)

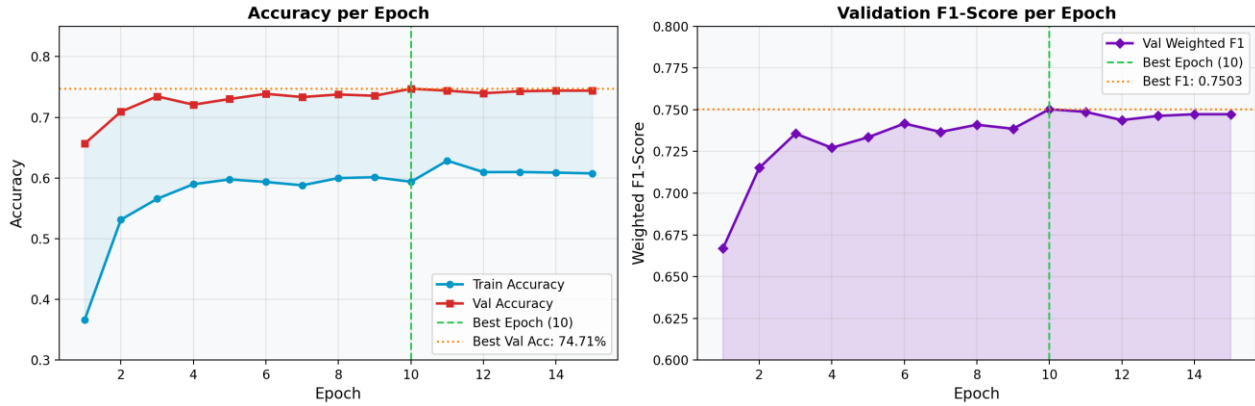


Fig. 6. Training curves on MVSA-Multiple. Best checkpoint at epoch 10 (Val. Acc.: 74.71%, Val. F1: 0.7503).

Per-Class F1-Score Progression — MVSA-Multiple

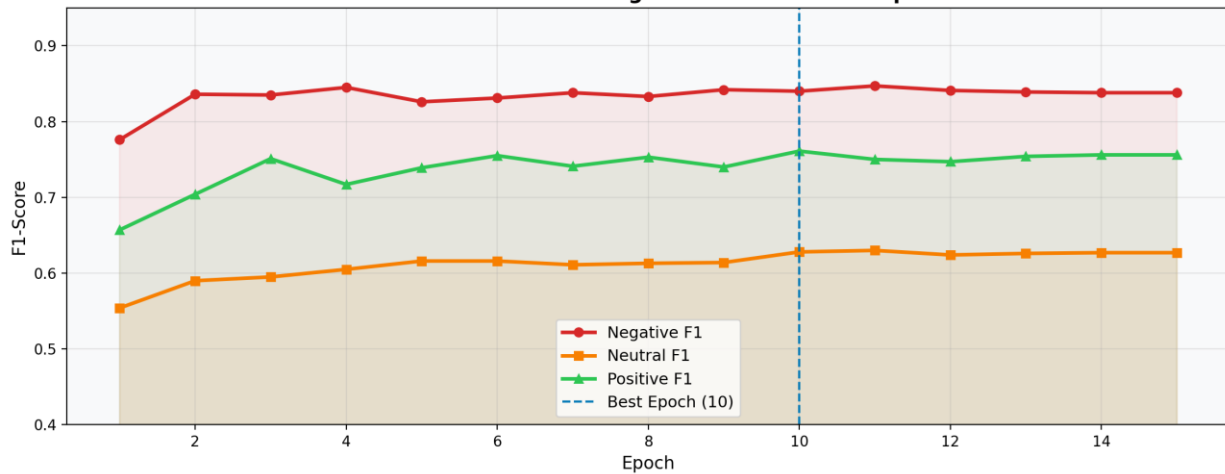


Fig. 7. Per-class F1 progression on MVSA-Multiple. The neutral class improves from 0.59 to 0.63 with targeted class weighting.

E. Training Dynamics on MVSA-Single

Fig. 8 and Fig. 9 present the training dynamics on MVSA-Single for the base model. Fig. 8 shows validation accuracy stabilising around epoch 17, with the best checkpoint at

validation accuracy 73.61% and validation F1 0.7262. Fig. 9 tracks per-class F1 progression and highlights the delayed emergence of the Neutral class, which starts from 0.00 at epoch 1 due to the severe class imbalance of only 10.4% Neutral samples.

CLIP-CrossFusion Net — Training Curves (MVSA-Single)

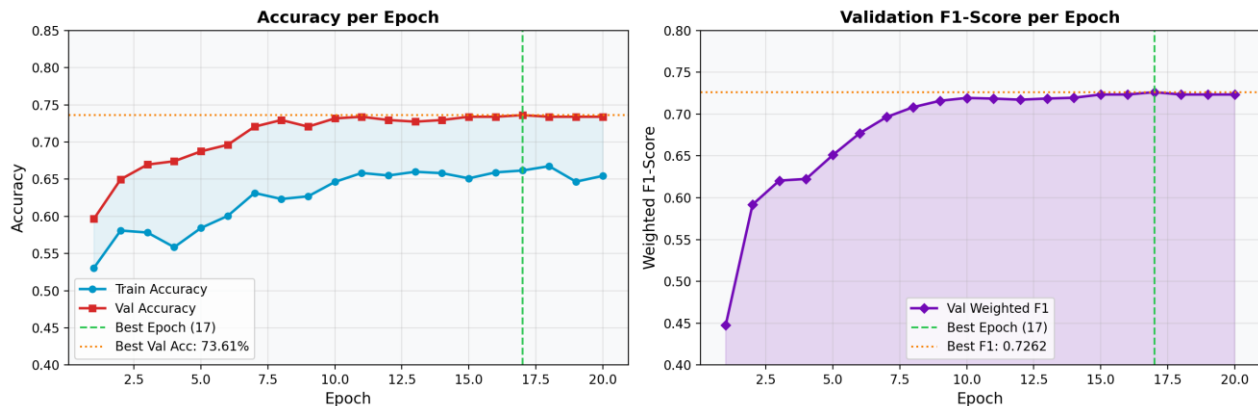


Fig. 8. Training curves on MVSA-Single (base model). Best checkpoint at epoch 17 (Val. Acc.: 73.61%, Val. F1: 0.7262).

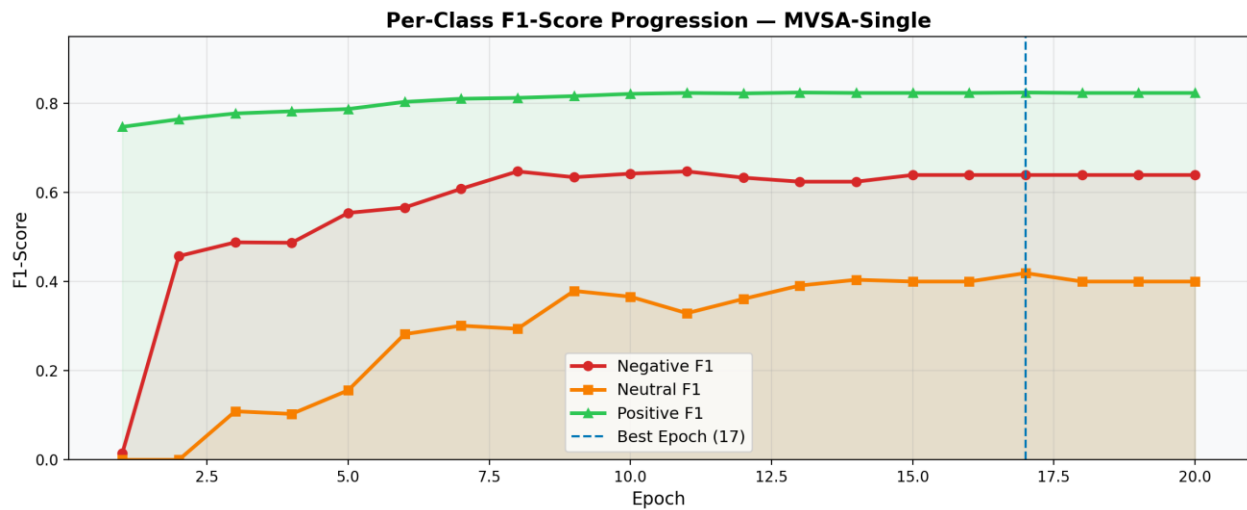


Fig. 9. Per-class F1 progression on MVSA-Single. The neutral class starts from 0 at epoch 1 due to severe class imbalance (10.4% of samples).

TABLE VI. DETAILED PER-CLASS RESULTS ON BOTH DATASETS (TEST SET WITH TTA)

Dataset / Model	Class	Prec.	Rec.	F1	Supp.	Observation
MVSA-Multiple, Focal + CW (74.39%)	Negative	0.91	0.80	0.85	893	Highest precision; clear negative signals
	Neutral	0.57	0.71	0.63	759	Hardest class; boosted by class weighting
	Positive	0.77	0.73	0.75	1,195	Largest class; good balance
MVSA-Single, ensemble (75.66%)	Negative	0.70	0.64	0.67	136	Stronger with ensemble than base alone
	Neutral	0.49	0.43	0.45	47	+0.07 over base alone; very small support
	Positive	0.82	0.87	0.85	269	Best among all versions

Table VI presents the per-class results on the MVSA-Single and MVSA-Multiple datasets. From Table VI, it is observed that, for the neutral class, only 47 samples are available as support.

F. Confusion Matrix Analysis

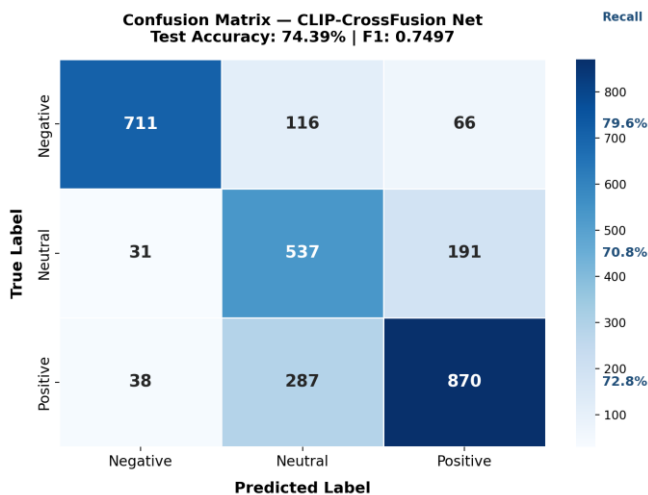


Fig. 10. Confusion matrix on the MVSA-Multiple test set (2,847 samples). Neutral to positive (191 cases) is the most frequent error.

Fig. 10 and Fig. 11 present confusion matrices for MVSA-Multiple and MVSA-Single, respectively. On MVSA-Multiple (Fig. 10, 2,847 samples), the most frequent error is Neutral

samples being misclassified as Positive (191 cases), reflecting the contextual ambiguity of the Neutral class. On MVSA-Single (Fig. 11, 452 samples, base model), the Neutral class remains the hardest, with reliable estimation further limited by a test support of only 47 Neutral samples.

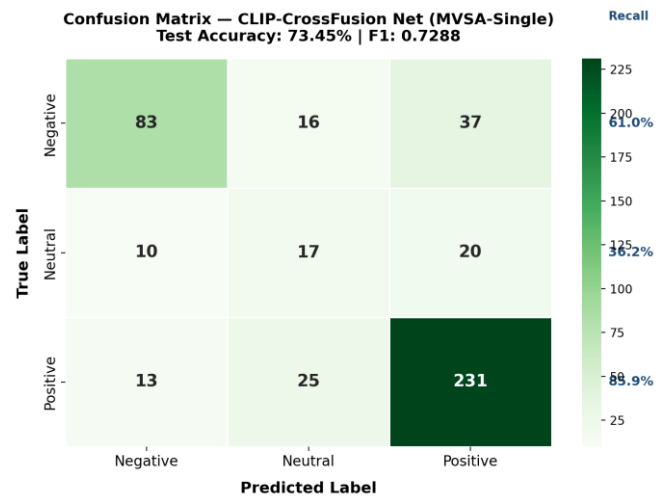


Fig. 11. Confusion matrix on the MVSA-Single test set (base model, 452 samples). Neutral is the most challenging class with only 47 test samples.

G. Cross-Dataset Comparison

Fig. 12 presents the final comparison across both datasets. The MVSA-Single ensemble slightly outperforms MVSA-

Multiple on overall accuracy (75.66% vs 74.39%) and Positive F1 (0.85 vs 0.75), while MVSA-Multiple is stronger on Negative F1 (0.85 vs 0.67) and Neutral F1 (0.63 vs 0.45). The Neutral gap reflects the much smaller Neutral support on

MVSA-Single (10.4% vs 26.7%). MVSA-Multiple leads on Negative and Neutral F1 due to larger, more balanced class support.

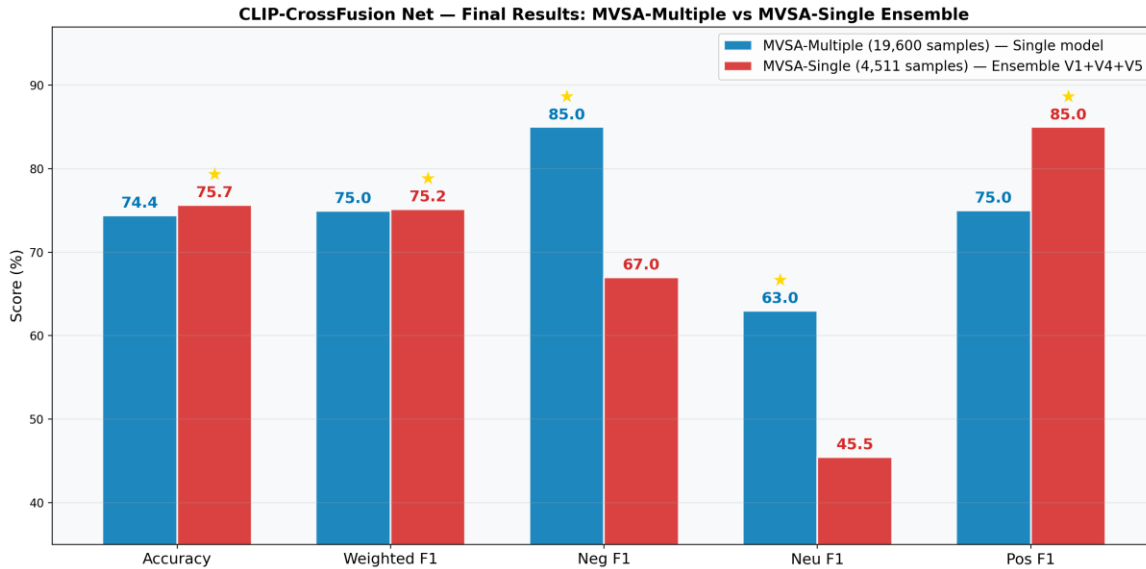


Fig. 12. Cross-dataset comparison of the final best models.

H. Comparison with State-of-the-Art

On MVSA-Multiple, the proposed model attains 74.39%, above ICCI (73.29%) and well above the concatenation baselines (56-58%). On MVSA-Single, the proposed ensemble reaches 75.66%, a 3.67 percentage-point gap from ICCI (79.33%). ICCI uses HSV colour features, which motivated the colour feature branch. The remaining gap likely comes from the richer colour modelling strategy of ICCI and different training data preprocessing. Biswas et al. [15] report 93.48% on MVSA-Multiple but use a highly specialised multi-stage pipeline with Mask-RCNN, AdaBoost, and Dempster-Shafer fusion that substantially increases complexity. The proposed end-to-end framework achieves competitive results with a simpler, more reproducible approach.

VII. CONCLUSION

This study presents CLIP-CrossFusion Net, a multimodal sentiment analysis framework combining CLIP ViT-B/32 and RoBERTa through bidirectional cross-modal attention for 3-class sentiment classification. The proposed model is evaluated comprehensively on both the MVSA-Multiple and MVSA-Single datasets with four model variants and a multi-model ensemble.

The main contributions and findings of the study are: 1) the proposed bidirectional cross-modal attention improves over simple concatenation by 12-15 percentage points on both datasets; 2) focal loss with class-balanced weighting pushes Neutral F1 from 0.59 to 0.63 on MVSA-Multiple and from 0.00 to 0.38 on MVSA-Single; 3) HSV colour features in the colour-contrastive variant add 0.67 percentage points of accuracy and improve Neutral F1 on MVSA-Single; 4) the multi-model ensemble achieves 75.66% on MVSA-Single, 2.21 percentage points over the single best model; and 5) both datasets reach

above 74% accuracy on a single consumer GPU, demonstrating practical feasibility.

In future, the following works may be explored: region-based visual attention for fine-grained sentiment localisation; caption generation for image-to-text conversion as used by ICCI; multi-task learning with modality alignment objectives; larger vision-language models (CLIP ViT-L/14 with memory-efficient training); and extension to video-based multimodal sentiment analysis.

DECLARATIONS

Conflict of Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Author Contributions. T S Kaveri and C K Roopa: conceptualization, methodology, investigation, writing - original draft; B S Harish and M S Kendagannaswamy: data curation, formal analysis, visualization; B S Harish: supervision, writing - review and editing.

Funding. The authors did not receive support from any organization for the submitted work.

Clinical Trial Registration. Not applicable. This study did not involve a clinical trial; therefore, no clinical trial registration number is available.

Consent to Publish. Not applicable.

Ethical Statement. This research did not involve human participants, animals, or clinical data. The study was conducted using publicly available datasets and computational machine learning methods. No ethical approval or informed consent was

required. The authors complied with all applicable institutional and journal ethical guidelines.

Consent to Participate. Not applicable; no human participants were involved.

Data Availability Statement. The datasets analysed during this study are publicly available from the repositories cited in the manuscript. All relevant dataset sources, accession details, and URLs are provided in the references and/or the dataset section (Section III). No new datasets were generated during this study.

REFERENCES

- [1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98-125, 2017.
- [2] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *MultiMedia Modeling*, Cham, Switzerland: Springer, 2016, pp. 15-27.
- [3] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 417-424.
- [4] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*, 2014, pp. 1746-1751.
- [5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, 2019, pp. 4171-4186.
- [8] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
- [9] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169-176.
- [10] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 328-339.
- [11] T. Zhu *et al.*, "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, vol. 25, pp. 3375-3385, 2023.
- [12] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings Assoc. Comput. Linguistics: NAACL 2022*, 2022, pp. 2282-2294.
- [13] G. Mu, Y. Chen, X. Li, L. Dai, and J. Dai, "Semantic enhancement and cross-modal interaction fusion for sentiment analysis in social media," *PLOS ONE*, vol. 20, no. 4, p. e0321011, 2025.
- [14] J. An and W. M. N. W. Zainon, "Integrating color cues to improve multimodal sentiment analysis in social media," *Eng. Appl. Artif. Intell.*, vol. 126, p. 106874, 2023.
- [15] S. Biswas, K. Young, and J. Griffith, "Exploring multimodal features for sentiment classification of social media data," Singapore: Springer, 2024, pp. 527-537, doi: 10.1007/978-981-99-8324-7_44.
- [16] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748-8763.
- [17] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [18] H. Touvron *et al.*, "Training data-efficient image transformers and distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10347-10357.
- [19] Z. Zhang, W. Wang, L. Zhu, and Z. Tang, "TAG-fusion: Two-stage attention guided multi-modal fusion network for semantic segmentation," *Digit. Signal Process.*, vol. 156, p. 104807, 2025.
- [20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.