

UniMSE: a unified approach for multimodal sentiment analysis leveraging the CMU-MOSI Dataset

Miriyala Trinath Basu, Mainak Saha, Arpita Gupta, Sumit Hazra, Shahin Fatima,
Chundakath House Sumalakshmi, Nallagopu Shanvi, Nyalapatla Anush Reddy,
Nallamalli Venkat Abhinav, Koganti Hemanth

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India

Article Info

Article history:

Received Dec 7, 2024

Revised Apr 22, 2025

Accepted Jul 3, 2025

Keywords:

Data fusion

Emotion recognition

MOSI

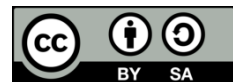
Sentiment analyze

UniMSE

ABSTRACT

This paper explores multimodal sentiment analysis using the CMU-MOSI dataset to enhance emotion detection through a unified approach called UniMSE. Traditional sentiment analysis, often reliant on single modalities such as text, faces limitations in capturing complex emotional nuances. UniMSE overcomes these challenges by integrating text, audio, and visual cues, significantly improving sentiment classification accuracy. The study reviews key datasets and compares leading models, showcasing the strengths of multimodal approaches. UniMSE leverages task formalization, pre-trained modality fusion, and multimodal contrastive learning, achieving superior performance on widely used benchmarks like MOSI and MOSEI. Additionally, the paper addresses the difficulties in effectively fusing diverse modalities and interpreting non-verbal signals, including sarcasm and tone. Future research directions are proposed to further advance multimodal sentiment analysis, with potential applications in areas like social media monitoring and mental health assessment. This work highlights UniMSE's contribution to developing more empathetic artificial intelligence (AI) systems capable of understanding complex emotional expressions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Miriyala Trinath Basu

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation

Hyderabad, 500075, Telangana, India

Email: tmiriyala@gmail.com

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, is a fundamental task in natural language processing (NLP) that aims to identify and classify subjective information such as opinions, emotions, and attitudes within text or speech [1]. Traditional sentiment analysis methods have largely relied on unimodal data, primarily textual input, using lexical features or deep learning-based approaches. While these techniques have proven effective in many applications, they often fail to capture the non-verbal cues that are critical to understanding the full spectrum of human emotion [2]-[4].

In real-world communication, people express emotions not just through words, but also through intonation, facial expressions, and gestures. These complementary signals provide essential context that can either reinforce or contradict spoken language [5]. To address this gap, multimodal sentiment analysis (MSA) has emerged as a research direction that integrates multiple modalities typically text, audio, and visual to better capture the emotional intent behind human expressions [6]. This has significant implications for areas such as emotion-aware virtual assistants, social robotics, and mental health analysis.

However, MSA comes with several challenges. First, the heterogeneous nature of modalities introduces difficulties in temporal alignment, feature integration, and signal imbalance [7]. Second, datasets often vary in labeling granularity and sentiment definitions, making it difficult to develop models that generalize well across tasks [8]. Third, ambiguous or conflicting emotional signals, such as sarcasm or subtle expressions, can reduce model reliability [9]. Although recent models have made progress using attention mechanisms, transformers, and memory networks, many remain task-specific and struggle with generalization and robustness [10].

To overcome these issues, this paper proposes UniMSE, a unified framework for multimodal sentiment analysis. UniMSE introduces task formalization to harmonize diverse label schemes, enabling consistent cross-dataset training [11]. It leverages modality-specific encoders for feature extraction and applies a hybrid fusion strategy that combines early and late fusion for flexible integration. Furthermore, an inter-modal contrastive learning objective is introduced to align modality representations and enhance performance under ambiguous or noisy input conditions [12], [13].

The key contributions of this paper are threefold:

- First, we propose a unified deep learning-based architecture that systematically integrates three modalities.
- Second, we develop a contrastive learning mechanism that reinforces modality alignment while improving generalizability.
- Third, we validate the model across four benchmark datasets CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP demonstrating superior performance in both accuracy and F1-score compared to state-of-the-art models.

This research directly addresses the challenges outlined above and shows how the proposed framework leads to measurable improvements, as demonstrated in the results. In addition to its empirical success, UniMSE presents opportunities for future application in emotion-aware systems, cross-lingual sentiment tasks, and human-computer interaction domains. Figure 1 illustrates the conceptual framework of UniMSE, highlighting the integration of text, audio, and visual modalities to provide a comprehensive understanding of sentiment through multimodal data fusion. The rest of the paper is structured as follows: section 2 reviews related work, section 3 presents the methodology, section 4 discusses results and analysis, and section 5 concludes the paper with key insights and future directions.

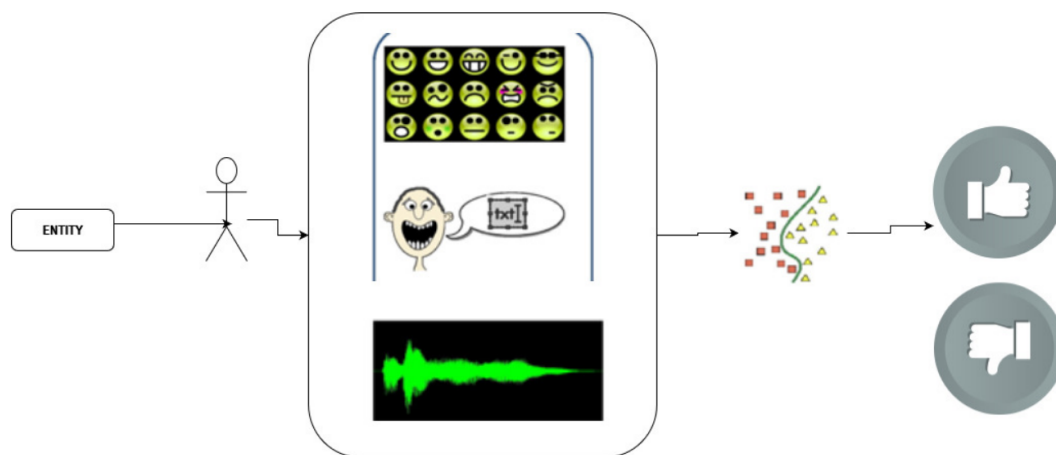


Figure 1. The image represents a conceptual framework for sentiment analysis using multimodal data, incorporating text, audio, and visual information to comprehensively assess sentiment

2. RELATED WORK

MSA has become increasingly important due to its ability to integrate verbal and non-verbal cues for better emotion recognition. Traditional sentiment analysis techniques were primarily text-based [14], and while effective for extracting lexical sentiment, they often overlooked critical paralinguistic elements such as vocal tone, facial expression, and gesture. These shortcomings motivated researchers to explore multimodal approaches that fuse text, audio, and visual data for improved sentiment classification.

To this end, several models have emerged with varying fusion strategies and architecture designs. Liu *et al.* [15] introduced the aspect-based attention and fusion network (ABAFN), employing attention mechanisms to highlight salient contextual and visual features for improved classification. Yu *et al.* [16]

proposed the image-target matching (ITM) network, which aligns image and text features through a coarse-to-fine mechanism and utilizes transformer layers for deep multimodal fusion.

Other studies have explored memory networks and graph-based methods. The multi-interactive memory network (MIMN) by Xu *et al.* [17] utilizes dual memory modules to model intra-modal and cross-modal interactions. Zhao and Yang [18] developed the fusion with GCN and SE-ResNeXt (FGSN) framework, leveraging graph convolutional networks alongside attention mechanisms to fuse textual and visual representations. Similarly, Wang *et al.* [19] proposed the AMCGC model, which incorporates aspect-level co-attention and gated mechanisms for fine-grained sentiment detection.

Transformer-based models have also gained popularity. The hierarchical interactive multimodal transformer (HIMT) [20] and the hierarchical cross-modal transformer (HCT) [21] adopt self-attention and aspect-aware layers to capture intricate interdependencies across modalities. While these models demonstrate improved accuracy, they typically require extensive computational resources and often lack robustness in noisy or ambiguous data environments.

Despite these advancements, several limitations remain. Most existing models are designed with dataset-specific assumptions and struggle to generalize across diverse datasets due to inconsistent labeling schemes. Additionally, aligning information from heterogeneous modalities remains a challenge, particularly when conflicting signals arise, such as a positive utterance delivered with a sarcastic tone. Furthermore, many of these models rely on static or unimodal-focused fusion mechanisms that are inadequate for handling emotion ambiguity and subtle non-verbal cues [22], [23].

In response to these challenges, the proposed UniMSE framework introduces a unified and flexible approach to multimodal sentiment analysis. It incorporates task formalization to standardize label representations across datasets, enabling consistent training and evaluation. The model employs a hybrid fusion strategy that combines early and late fusion techniques, allowing it to adaptively integrate information from multiple modalities at different stages of processing. Additionally, inter-modal contrastive learning is applied to align semantically similar features while minimizing modality noise, thereby improving sentiment discrimination even under ambiguity. These innovations position UniMSE as a scalable and robust solution, capable of outperforming existing models in both accuracy and generalizability.

3. METHOD

The proposed UniMSE framework aims to enhance the robustness and generalizability of MSA by introducing a unified label formalization strategy, modality-specific feature extraction, hybrid fusion mechanisms, and inter-modal contrastive learning. This section provides a detailed explanation of the datasets, preprocessing pipeline, architectural components, and training setup.

3.1. Datasets

To evaluate the performance of UniMSE across diverse multimodal scenarios, we selected four benchmark datasets: CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP. These datasets cover a broad spectrum of contexts from YouTube monologues to television dialogues and dyadic conversations offering varying degrees of sentiment granularity, emotional intensity, and modality combinations. All four datasets include synchronized text, audio, and visual modalities. A detailed overview of these datasets, including their size, modality distribution, source platforms, and language, is presented in Table 1. The diversity of these benchmarks ensures a thorough evaluation of UniMSE's generalization capability across different domains and emotional contexts.

Table 1. Overview of popular multimodal sentiment datasets

Dataset	Year	Modalities	Size	Source	Language
IEMOCAP [24]	2008	Audio, video, text	1,039 segments	Speech Lab, USC	English
DEAP [25]	2011	EEG signals	32 participants	Queen Mary Univ.	English
MOSI [26]	2016	Audio, video, text	2,199 segments	YouTube	English
MOSEI [27]	2018	Audio, video, text	23,453 segments	YouTube	English
MELD [28]	2019	Audio, video, text	13,000 segments	Friends TV Series	English
Multi-ZOL [29]	2019	Text, video	5,288 reviews	ZOL.com	Chinese
CH-SIMS [30]	2020	Audio, video, text	2,281 clips	YouTube	Chinese
MOSEAS [31]	2021	Audio, video, text	40,000 segments	YouTube	Spanish, French
MALE-CALL [32]	2021	Audio, video, text	291 videos	YouTube	English
B-TASA [33]	2021	Text, video	4,700 tweets	Twitter	English
FACTIFY [34]	2022	Image, text	50,000 tweets	Twitter	English
MEMOTION [35]	2022	Image, text	10,000 memes	Reddit, Facebook	English

IEMOCAP [24] is a multimodal dataset featuring 1,039 conversational segments over 12 hours of video with various emotions expressed through audio and visual data. DEAP [25] focuses on physiological signals from EEG data recorded at 512 Hz from participants rating video stimuli on valence and arousal. MOSI [26] consists of 93 YouTube videos with sentiment intensity annotations across 2,199 opinion segments. MOSEI [27] expands on this with over 3,228 videos segmented into 23,453 parts across multiple modalities. MELD [28] includes video clips from the Friends TV series annotated for seven emotions and sentiment. Multi-ZOL [29] contains mobile phone reviews rated on a sentiment scale. CH-SIMS [30] is a Mandarin dataset focused on facial and voice data with sentiment intensity annotations. MOSEA [31] covers multilingual sentence fragments with sentiment ratings. MALE-CALL [32] consists of YouTube videos in English across multiple modalities. B-TASA [33] includes tweets combining video and text for social media sentiment analysis. FACTIFY [34] is aimed at fake news detection with image-text data points categorized into support or refutation claims. Finally, MEMOTION [35] provides memes annotated for sentiment categories like humor and sarcasm. These datasets are crucial for advancing multimodal sentiment analysis research by offering diverse sources of data that enhance sentiment detection across various contexts.

3.2. Overview of the UniMSE framework

The UniMSE framework is built to unify the learning process across sentiment and emotion classification tasks while addressing challenges such as modality misalignment, label inconsistency, and sentiment ambiguity [36]. It incorporates four key components: modality-specific feature extraction, unified task formalization, a hybrid fusion strategy, and inter-modal contrastive learning.

The system processes raw text, audio, and visual data using dedicated feature extractors, including LSTM-based models for audio (A-LSTM) and visual (V-LSTM) streams, and a T5 encoder for textual input. These extracted embeddings are aligned and fed into a multimodal fusion layer within a transformer-based encoder-decoder structure. A contrastive learning loss is applied to ensure semantically similar embeddings across modalities are brought closer, while dissimilar ones are pushed apart. This mechanism improves representation robustness, especially in emotionally ambiguous scenarios [37].

Figure 2 illustrates the conceptual architecture of UniMSE, showing the end-to-end flow from modality-specific inputs to fused embeddings and task-specific outputs. The framework supports both MSA and emotion recognition in conversation through a shared architecture and decoding layer.

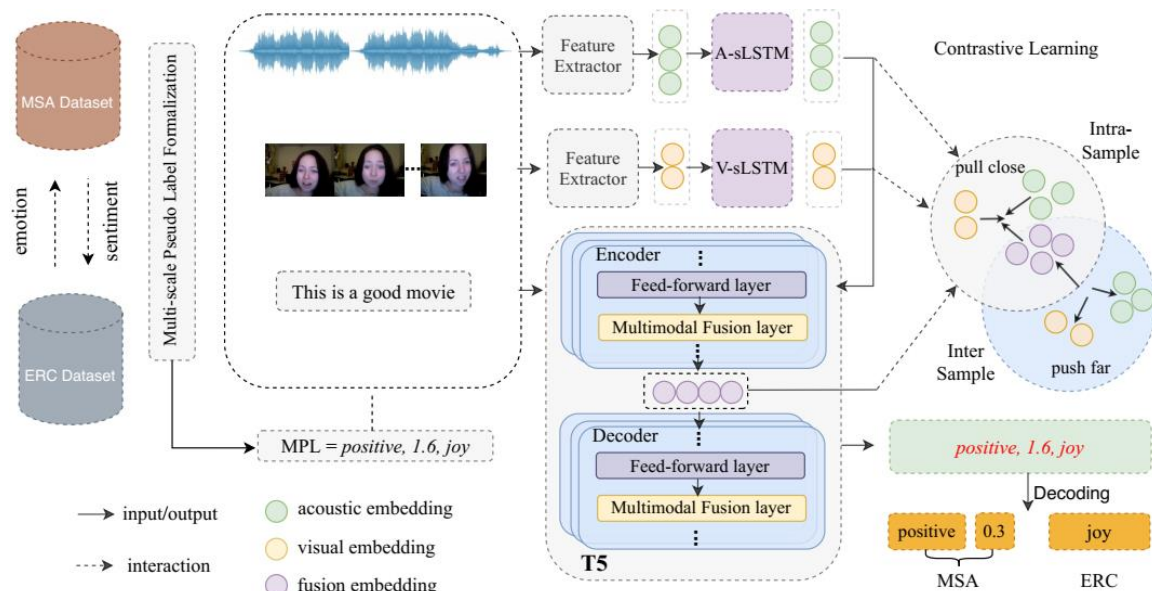


Figure 2. Conceptual framework of the proposed UniMSE model, integrating text, audio, and visual modalities with fusion and contrastive learning for sentiment and emotion classification

3.3. Data preprocessing

Each modality underwent preprocessing tailored to its nature. Text data was cleaned through lowercasing, stopword removal, and punctuation normalization. Utterances were tokenized using a pre-trained T5-compatible tokenizer to retain contextual richness. For audio, acoustic features such as mel-frequency cepstral coefficients (MFCCs), pitch, and energy were extracted using the openSMILE toolkit.

These features were normalized and temporally aligned with corresponding visual frames. Visual preprocessing involved sampling frames from videos at 5–10 FPS and extracting facial landmarks and expressions using pre-trained CNNs. All modalities were aligned using timestamp information to ensure consistent cross-modal representation.

3.4. Task formalization

Given that sentiment datasets differ in annotation schemes ranging from binary sentiment labels to fine-grained continuous scores a task formalization step was required. For datasets like CMU-MOSI and MOSEI, which contain sentiment scores from -3 to +3, values were discretized into three classes: negative, neutral, and positive, using empirically defined thresholds. For datasets with pre-existing categorical labels, a direct mapping to this unified three-class framework was applied. This formalization allowed UniMSE to learn a consistent sentiment classification task across heterogeneous datasets, facilitating transfer learning and standardized evaluation.

3.5. Feature extraction

UniMSE employs modality-specific neural encoders to extract discriminative representations. Text features are encoded using a pre-trained T5 encoder, capturing both semantic and contextual dependencies. Audio features, which exhibit strong temporal characteristics, are processed using bidirectional long short-term memory networks to model patterns such as intonation and rhythm [38]. Visual features are extracted using convolutional neural networks, focusing on facial expressions and micro-gestures. All extracted embeddings are projected into a shared latent space to support effective multimodal fusion.

3.6. Hybrid fusion strategy

To address the limitations of purely early or late fusion, UniMSE integrates a hybrid fusion mechanism. In the early stage, feature vectors from each modality are concatenated and passed through a self-attention layer to capture fine-grained inter-modal dependencies. This is followed by a decision-level fusion, where intermediate predictions from each modality stream are reweighted using a learnable gating mechanism, allowing the model to dynamically emphasize or de-emphasize modality contributions based on context [39]. This strategy enhances interpretability and robustness, especially in the presence of noisy or missing modalities.

3.7. Inter-modal contrastive learning

To further enhance cross-modal alignment and mitigate representation noise, UniMSE incorporates inter-modal contrastive learning. This training objective encourages embeddings from different modalities but similar sentiment classes to converge in the shared space while pushing apart dissimilar samples. The contrastive loss is computed using cosine similarity and a temperature-scaled SoftMax function:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(M_i, M_j)/\tau)}{\sum_{k=1}^N \exp(\text{sim}((M_i, M_k)/\tau)} \quad (1)$$

In (1), $\mathcal{L}_{contrastive}$ denotes the contrastive loss, which is minimized during training to improve inter-modal consistency. The term $\text{sim}(M_i, M_j)$ represents the cosine similarity between modality embeddings M_i and M_j belonging to the same sentiment class. The denominator includes all M_k , representing other embeddings in the batch, used to normalize the similarity score. The variable τ is a temperature scaling parameter that controls the sharpness of the SoftMax distribution, and N is the number of samples in the batch [40].

3.8. Training configuration

UniMSE is trained using a composite loss function that includes categorical cross-entropy for sentiment classification and contrastive loss for inter-modal alignment. The model is optimized using the Adam optimizer with a learning rate of 1e-4 and a batch size of 32. Early stopping based on validation loss is used to prevent overfitting. All experiments are conducted using the PyTorch framework on systems equipped with NVIDIA GPUs. Hyperparameters are tuned via grid search, using validation accuracy and F1-score as the main evaluation criteria.

4. RESULTS AND DISCUSSION

To evaluate the performance of the proposed UniMSE framework, we conducted extensive experiments on four benchmark datasets: CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP. Each dataset presents unique challenges ranging from sentiment granularity and emotional ambiguity to modality noise

and conversational variability making them ideal for assessing the robustness and generalizability of multimodal models. Evaluation was performed using widely accepted metrics, including mean absolute error (MAE), correlation (Corr), binary and 7-class accuracy (ACC-2, ACC-7), and macro F1-score for CMU-MOSI and MOSEI, and accuracy (ACC) and weighted F1-score (WF1) for MELD and IEMOCAP.

Across all four datasets, UniMSE consistently outperformed state-of-the-art models. On CMU-MOSI, UniMSE achieved a MAE of 0.691, Corr of 0.839, ACC-2 of 85.48%, and an F1-score of 85.90%, surpassing strong baselines such as MAG-BERT, Self-MM, and MulT. On CMU-MOSEI, which includes more than 23,000 opinion segments, UniMSE attained a MAE of 0.523, Corr of 0.773, and a 7-class accuracy of 85.86%, with an F1-score of 87.97%, setting a new benchmark in multimodal sentiment prediction.

In the emotion-oriented MELD and IEMOCAP datasets, which emphasize conversational context, UniMSE also demonstrated significant improvements. For MELD, the model achieved an accuracy of 65.51% and WF1 of 65.91%, outperforming previous models such as MMIM and MMGCN. On IEMOCAP, UniMSE reached 70.66% accuracy and a WF1 of 70.46%, outperforming models including DialogXL and COSMIC.

These results highlight the effectiveness of UniMSE's hybrid fusion strategy and inter-modal contrastive learning. Unlike traditional early or late fusion methods, UniMSE dynamically adapts the contribution of each modality, allowing it to better capture the nuances of sentiment expression. The contrastive loss, in particular, was beneficial for emotionally ambiguous samples—such as sarcastic or conflicting utterances by enhancing alignment between semantically similar signals across modalities. A comparative performance analysis with leading models is presented in Table 2. The table summarizes UniMSE's consistent improvements across all tasks, with especially notable gains in macro F1-score a critical metric for evaluating performance on imbalanced datasets such as MELD.

Table 2. Comparative performance of UniMSE and existing state-of-the-art models across CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP datasets

Method	MOSI (MAE↓)	MOSI (Corr↑)	MOSI (ACC -2↑)	MOSI (ACC -7↑)	MOS I (F1↑)	MOSEI (MAE↓)	MOSE I (Corr↑)	MOSE I (ACC- 2↑)	MOSE I (ACC- 7↑)	MOSE I (F1↑)	MELD (ACC↑)	MELD (WF1↑)	IEMOCA P (ACC↑)	IEMOCA P (WF1↑)
LMF	0.917	0.655	82.5	-	-	0.633	0.703	82	-	-	-	-	-	-
TFM	0.907	0.696	84.3	-	-	0.614	0.723	83.7	-	-	-	-	-	-
MFM	0.836	0.714	84.4	-	-	0.561	0.733	84.2	-	-	-	-	-	-
MTAG	0.906	0.711	84.3	-	-	0.651	0.735	83.7	-	-	-	-	-	-
SPC	0.812	0.714	84.1	-	-	0.51	0.74	84.6	-	-	-	-	-	-
ICCN	0.738	0.731	84.2	-	-	0.544	0.756	84.2	-	-	-	-	-	-
MuT	0.871	0.714	84.4	-	-	0.565	0.75	85.1	-	-	-	-	-	-
MSA	0.804	0.744	85.02	84.83	85.07	0.544	0.756	82.9	84.99	83.46	-	-	-	-
Self-MM	0.713	0.751	84.14	85.23	84.12	0.51	0.748	84.15	85.99	84.27	-	-	-	-
MAG- BERT	0.712	0.761	84.25	85.12	84.93	0.517	0.754	82.72	84.99	83.46	-	-	-	-
MMIM	-	-	-	-	-	-	-	-	-	-	59.46	59.49	65.1	65.38
DialogGCN	-	-	-	-	-	-	-	-	-	-	-	-	64.18	63.54
DialogXL	-	-	-	-	-	-	-	-	-	-	-	-	66.35	66.53
EMRG- DialogGCN	-	-	-	-	-	-	-	-	-	-	-	-	64.6	64.74
COG- BART	-	-	-	-	-	-	-	-	-	-	-	-	65.63	66.1
Psychologis t	-	-	-	-	-	-	-	-	-	-	-	-	65.72	66.22
COSMIC	-	-	-	-	-	-	-	-	-	-	-	-	66.24	66.39
TODKAV	-	-	-	-	-	-	-	-	-	-	-	-	64.12	64.4
MMGCN	-	-	-	-	-	-	-	-	-	-	-	-	65.66	65.91
MM-DFN	-	-	-	-	-	-	-	-	-	-	-	-	64.44	64.72
UniMSE	0.691	0.839	85.48	85.56	85.9	0.523	0.773	84.23	85.86	87.97	65.51	65.91	70.66	70.46

Despite its strengths, UniMSE has limitations. The model's performance degraded slightly in scenarios involving extreme modality imbalance, such as missing visual cues or noisy audio input. This suggests a need for further robustness techniques, such as modality dropout or uncertainty-aware fusion. Additionally, the model's decision-making process remains a black box; while attention mechanisms improve adaptivity, a deeper understanding of why certain modalities are prioritized remains unexplored. Future work could incorporate explainability tools, such as saliency mapping or attention visualization, to improve interpretability for deployment in high-stakes settings. Overall, the experimental results confirm that UniMSE successfully addresses core challenges in multimodal sentiment analysis. Its strong performance, generalizability, and resilience to ambiguity position it as a viable solution for real-world applications, including emotion-aware artificial intelligence (AI) systems, human-computer interaction, and social media analysis.

5. CONCLUSION

This paper introduced UniMSE, a unified framework for multimodal sentiment analysis that integrates text, audio, and visual modalities through task formalization, hybrid fusion, and inter-modal contrastive learning. The model was developed to overcome common limitations in the field, such as inconsistent task definitions, weak modality alignment, and reduced performance in ambiguous emotional contexts. To address these, UniMSE introduced a unified label mapping scheme, leveraged modality-specific encoders, and employed a dual-stage fusion strategy supported by a contrastive learning objective.

Extensive experiments on four benchmark datasets CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP demonstrated that UniMSE outperforms existing models in both accuracy and F1-score. Its consistent performance across datasets validates its generalizability and effectiveness in handling diverse sentiment and emotion classification scenarios. Particularly, the hybrid fusion mechanism allowed for dynamic modality weighting, while contrastive learning improved robustness in cases of sentiment ambiguity or modality conflict.

The results obtained in this study validate the objectives outlined in the introduction. The unified task formalization, hybrid fusion strategy, and contrastive learning components proposed in UniMSE were empirically shown to address the key challenges of multimodal sentiment analysis, including cross-modal alignment, task inconsistency, and emotion ambiguity. This direct alignment between the research goals and the achieved outcomes confirms the methodological soundness and practical value of the approach. Furthermore, the adaptability of UniMSE across datasets suggests promising potential for future applications in emotion-aware systems, social media monitoring, human-computer interaction, and multilingual sentiment environments.

Nevertheless, several limitations remain. The model’s performance may be affected in cases of severely corrupted or missing modalities, and its internal decision-making process lacks full interpretability. Future work may focus on integrating adaptive modality dropout, attention visualization, or explainability modules. Additionally, extending the framework to support multilingual sentiment recognition and domain-specific tuning (e.g., medical, educational, or low-resource languages) presents valuable directions for further research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Miriyala Trinath Basu	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mainak Saha	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Arpita Gupta						✓	✓					✓		✓
Sumit Hazra						✓	✓					✓		✓
Shahin Fatima												✓		✓
Chundakath House												✓		✓
Sumalakshmi														
Nallagopu Shanvi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Nyalapatla Anush Reddy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Nallamalli Venkat Abhinav	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Koganti Hemanth	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				

C : C onceptualization	I : I nterpretation	Vi : V isualization
M : M ethodology	R : R esources	Su : S upervision
So : S oftware	D : D ata Curation	P : P roject administration
Va : V alidation	O : Writing - O riginal Draft	Fu : F unding acquisition
Fo : F ormal analysis	E : Writing - Review & E ditng	

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known financial, personal, or professional conflicts of interest that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are publicly available from third-party sources. Specifically, datasets such as CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP were obtained from their respective official repositories. No new data were created or collected by the authors for this study. All analysis was conducted using these openly accessible resources.




REFERENCES

- [1] R. S. Kumar, A. F. Saviour Devaraj, M. Rajeswari, E. G. Julie, Y. H. Robinson, and V. Shanmuganathan, "Exploration of sentiment analysis and legitimate artistry for opinion mining," *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11989–12004, Jan. 2022, doi: 10.1007/s11042-020-10480-w.
- [2] R. Das and T. D. Singh, "Multimodal sentiment analysis: a survey of methods, trends, and challenges," *ACM Computing Surveys*, vol. 55, no. 13, pp. 1–38, Jul. 2023, doi: 10.1145/3586075.
- [3] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3375–3385, 2023, doi: 10.1109/TMM.2022.3160060.
- [4] A. Anikin, "The link between auditory salience and emotion intensity," *Cognition and Emotion*, vol. 34, no. 6, pp. 1246–1259, Mar. 2020, doi: 10.1080/02699931.2020.1736992.
- [5] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861–881, Jan. 2020, doi: 10.1111/coin.12274.
- [6] S. M. Kim and E. Hovy, "Determining the sentiment of opinions," in *COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 1367-es, doi: 10.3115/1220355.1220555.
- [7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, Mar. 2013, doi: 10.1109/MIS.2013.30.
- [8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, Jan. 2023, doi: 10.1145/3560815.
- [9] N. Li, C. Y. Chow, and J. D. Zhang, "SEML: a semi-supervised multi-task learning framework for aspect-based sentiment analysis," *IEEE Access*, vol. 8, pp. 189287–189297, 2020, doi: 10.1109/ACCESS.2020.3031665.
- [10] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: a comparative review," *Expert Systems with Applications*, vol. 118, pp. 272–299, Mar. 2019, doi: 10.1016/j.eswa.2018.10.003.
- [11] J. Li *et al.*, "Hybrid multimodal feature extraction, mining and fusion for sentiment analysis," in *MuSe 2022 - Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge*, Oct. 2022, pp. 81–88, doi: 10.1145/3551876.3554809.
- [12] A. Favaro *et al.*, "A multi-modal array of interpretable features to evaluate language and speech patterns in different neurological disorders," in *2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings*, Jan. 2023, pp. 532–539, doi: 10.1109/SLT54892.2023.10022435.
- [13] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.
- [14] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.
- [15] L. L. Liu, Y. Yang, and J. Wang, "ABAFN: aspect-based sentiment analysis model for multimodal," *Computer Engineering and Applications Journal (ComEngApp)*, vol. 58, no. 10, pp. 193–199, 2022.
- [16] J. Yu, J. Wang, R. Xia, and J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching," in *IJCAI International Joint Conference on Artificial Intelligence*, Jul. 2022, pp. 4482–4488, doi: 10.24963/ijcai.2022/622.
- [17] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 33, no. 01, pp. 371–378, Jul. 2019, doi: 10.1609/aaai.v33i01.3301371.
- [18] J. Zhao and F. Yang, "Fusion with GCN and SE-ResNeXt network for aspect based multimodal sentiment analysis," in *ITNEC 2023 - IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference*, Feb. 2023, pp. 336–340, doi: 10.1109/ITNEC56291.2023.10082618.
- [19] S. J. Wang, G. Y. Cai, G. R. Lv, and W. B. Tang, "Aspect-level multimodal co-attention graph convolutional sentiment analysis model," *International Journal of Image and Graphics*, pp. 1–16, 2023.
- [20] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1966–1978, Jul. 2023, doi: 10.1109/TAFFC.2022.3171091.
- [21] K. Chen, X. G. Dong, and X. S. Zhou, "Research on multimodal fine-grained sentiment analysis method based on cross-modal transformer," *Computer and Digital Engineering*, vol. 50, no. 10, pp. 2270–2275, 2022.
- [22] S. Munirathinam, "Industry 4.0: industrial internet of things (IIoT)," in *Advances in Computers*, vol. 117, no. 1, Elsevier, 2020, pp. 129–164.
- [23] K. Zhao, M. Zheng, Q. Li and J. Liu, "Multimodal sentiment analysis—a comprehensive survey from a fusion methods perspective," in *IEEE Access*, vol. 13, pp. 64556–64583, 2025, doi: 10.1109/ACCESS.2025.3554665.
- [24] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.
- [25] S. Koelstra *et al.*, "DEAP: a database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.
- [26] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
- [27] A. Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, pp. 2236–2246, doi: 10.18653/v1/p18-1208.




- [28] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: a multimodal multi-party dataset for emotion recognition in conversations," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 527–536, doi: 10.18653/v1/p19-1050.
- [29] Z. Liu and J. Wang, "Scaling multimodal pre-training via cross-modality gradient alignment," *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.
- [30] W. Yu *et al.*, "CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3718–3727, doi: 10.18653/v1/2020.acl-main.343.
- [31] A. Zadeh, Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, and L. P. Morency, "CMU-MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 1801–1812, doi: 10.18653/v1/2020.emnlp-main.141.
- [32] Z. Li, S. Li, Y. Zhang, and S. Li, "MALE-CALL: a multimodal dataset for sentiment analysis of vehicle reviews," in *IEEE/ACM International Conference on Multimedia Conference (MM)*, 2021.
- [33] Y. Luo, J. Li, J. Chen, Y. Zhang, and X. Li, "B-TASA: a benchmark dataset for multimodal sentiment analysis on Twitter," in *IEEE/ACM International Conference on Multimedia Conference (MM)*, 2021.
- [34] S. Mishra *et al.*, "Factify: a multi-modal fact verification dataset.," in *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*, 2022.
- [35] S. Ramamoorthy *et al.*, "Memotion 2: Dataset on sentiment and emotion analysis of memes," in *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection*, CEUR, vol. 17, 2020.
- [36] G. Hu, T. E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022, pp. 7837–7851, doi: 10.18653/v1/2022.emnlp-main.534.
- [37] A. Zadeh, M. Chen, S. Poria, and L. P. Morency, "Multimodal sentiment analysis: a survey and research directions," *IEEE Transactions on Affective Computing*, 2018.
- [38] J. Li, Z. Zhang, B. Wang, Q. Zhao, and C. Zhang, "Inter-and intra-modal contrastive hybrid learning framework for multimodal abstractive summarization," *Entropy*, vol. 24, no. 6, p. 764, May 2022, doi: 10.3390/e24060764.
- [39] N. Wang and Q. Wang, "Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 1, pp. 1-19, 2024, doi: 10.1145/370299.
- [40] H. Wang, X. Li, Z. Ren, M. Wang and C. Ma, "Multimodal sentiment analysis representations learning via contrastive learning with condense attention fusion," *Sensors*, vol. 23, no. 5, p. 2679, 2023, doi: 10.3390/s23052679.

BIOGRAPHIES OF AUTHORS






Dr. Miriyala Trinath Basu    received his Ph.D. in engineering (Department of Computer science and Engineering) at K L University, Vijayawada, India. He is working as an Associate Professor in the Department of Computer Science and Engineering, K L Deemed to be University, Hyderabad. His research works have been published in numerous peer reviewed journals. He also has been an active reviewer for many peer reviewed journals. He can be contacted at email: tmiriyala@gmail.com.






Mainak Saha    is an Assistant Professor in the Department of Computer Science and Engineering at K L Deemed to be University, Hyderabad. He is currently pursuing a Ph.D. in computer science and engineering at the National Institute of Technology, Agartala. He earned his Master of Technology degree from Tripura University, Agartala, India. His research primarily focuses on AI, ML, image processing, and computer vision. His dedication and expertise in these fields contribute significantly to advancing research efforts. He can be contacted at email: mainak.skms@gmail.com.






Dr. Arpita Gupta    received her Ph.D. from NIT, Tiruchirappalli in Transfer Learning. She is working as an Associate Professor and HOD in the Department of Computer Science and Engineering, K L Deemed to be University, Hyderabad Aziz Nagar Campus. Her research works have been published in numerous peer reviewed journals. She also has been an active reviewer for many peer reviewed journals. She can be contacted at email: arpitagupta2993@gmail.com.






Dr. Sumit Hazra    received his Ph.D. degree in CSE specializing in AI and ML from NIT Rourkela, Odisha. He received his M. Tech degree in CSE from VIT University, Vellore, India. He is currently working as an Assistant Professor in Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Hyderabad. He has authored several journals, international and core-ranking conferences and book chapters. He is an active reviewer in many peer-reviewed journals. His research interests include gait analysis, cognitive science, AI, pattern recognition, and computer vision. He can be contacted at email: sumhaz15@gmail.com.






Dr. Shahin Fatima    is an Assistant Professor in the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Hyderabad. She has completed her Doctorate in computer science and engineering from Integral University, Lucknow, India. She has done her M. Tech and B. Tech in computer science and engineering. She has authored several journals, international conferences and book chapters. She is an active reviewer in many peer-reviewed journals and has total academic teaching experience of 10 years. Her research interests include cloud computing, data science, AI, pattern recognition, and blockchain. She can be contacted at email: shahin.fatima@klh.edu.in.






Dr. Chundakath House Sumalakshmi    is an Assistant Professor in the Department of Computer Science and Engineering at KL University, Hyderabad. She holds a Ph.D. in information and communication engineering from Anna University, specializing in facial expression recognition using machine learning and deep learning techniques. With over 13 years of teaching experience and a strong research background, she has published extensively in SCI and Scopus-indexed journals, focusing on artificial intelligence, image processing, and optimization techniques. Dr. Sumalakshmi has received multiple accolades, including the Innovative Teaching Excellence Award and the Best Woman Faculty Award. She is also a certified Advanced RPA Professional and an active mentor in skill development initiatives. She can be contacted at email: chundakath.sumalakshmi@klh.edu.in.







Nallagopu Shanvi    holds a Bachelor of Technology (B. Tech) degree in computer science and engineering from K L University, Hyderabad, with a specialization in data science. She has received an offer letter from Accenture and will be joining soon. She is dedicated to enhancing her technical skills and is eager to apply her academic knowledge in a professional environment, making her a promising future engineer. She can be contacted at email: shanvi1424@gmail.com.







Nyalapatla Anush Reddy    holds a Bachelor of Technology (B. Tech) degree in computer science and engineering from K L University, Hyderabad, with a specialization in data science. He is preparing to go abroad to pursue a master's degree in data science and artificial intelligence. He is committed to further honing his skills to achieve his professional aspirations. He can be contacted at email: reddymanush07@gmail.com.



Nallamalli Venkat Abhinav     holds a B. Tech in computer science and engineering, specializing in AI from KL University, Hyderabad, and is now pursuing a master's in IT and management at the University of Texas at Dallas. He has hands-on experience in market analysis and collaborating with cross-functional teams to deliver data-driven solutions. Abhinav thrives at the intersection of technology, user experience, and business strategy, aiming to apply these skills in a product management role. He can be contacted at email: anallamalli.1@gmail.com.



Koganti Hemanth     holds a Bachelor of Technology (B. Tech) degree in computer science and engineering from K L University, Hyderabad, with a specialization in data science. He is currently taking coaching for Java Full Stack. He will be going abroad in the spring intake to pursue a master's degree. He is focused on developing his skills further to achieve his career goals. He can be contacted at email: hemanthchowdarykoganti@gmail.com.