International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Multimodal Approach to Assessing Sentimental Intensity and Subjectivity using Unified MSE Model

Mohd Usman Khan[1], Faiyaz Ahamad[2]

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: In the dynamic realm of multimodal learning, where representation Learning serves as a pivotal key, our research introduces a groundbreaking approach to understanding sentiment and subjectivity in audio and text. Illustration from self-supervised learning, we've innovatively combined multi-modal and Unified--modal tasks, emphasizing the crucial aspects of consistency and distinctiveness. Our training techniques, likened to the art of fine-tuning an instrument, harmonize the learning process, prioritizing samples with distinctive supervisions. Addressing the pressing need for robust datasets and methodologies in combinational text and audio sentiment analysis, we offer the dataset for Multi-modal sentiment intensity assessment at the Opinion Level (MOSI). This meticulously annotated corpus offers insights into subjectivity, sentiment intensity, text features, and audio nuances, setting a benchmark for future research. Our method not only excels in generating Unified-modal supervisions but also stands resilient against benchmarks like MOSI and MOSEI, even competing human curated annotations on the challenging datasets. This pioneering work paves the way for deeper explorations and applications in the burgeoning field of sentiment analysis.

Keywords: *Multimodal Learning, Subjectivity Assessment, Audio & Text Analysis, Distinctiveness, Unified-modal Supervision.*

1. Introduction

The advancement of communication technologies and the emergence of social platforms such as Facebook and YouTube have resulted in the generation of a significant volume of multi-modal data infused with sentiment daily. Sentiment plays a crucial role in shaping human interactions and perceptions and has a profound impact on advancements in artificial intelligence. These advancements find applications in various domains, including human-machine dialogues and autonomous driving Chen et al.[1]iinducted a study proposing a method for multimodal sentiment analysis that involves word-level fusion and reinforcement learning. While text serves as a fundamental medium for expressing sentiments through words, phrases, and relationships, Li et al.[2] suggested enhancing bidirectional representations by employing context-aware embedding for aspect-based sentiment analysis. However, relying solely on text for sentiment analysis can sometimes present limitations. Perceptive emotions solely from text can be challenging, prompting real-world communication to frequently integrate audio cues with text. Audio modality captures sentiments through voice nuances like pitch, energy, and loudness Li et al. [3]aimed to enhance speech emotion recognition through discriminate representation learning. The synergy between text and audio modalities enhances sentiment analysis, providing a richer emotional context Majumder et al. [4] employed hierarchical fusion with context modeling for their multimodal sentiment analysis approach. For example, Figure 1 illustrates how the sentence “You’re aware he's the one who did it” can convey varied emotions based on its context. While the sentiment might seem ambiguous through text alone, the accompanying audio—perhaps a speaker's somber tone can clarify its negative connotation [5]. Recognizing the potential of combined modalities, multimodal sentiment analysis within affective computing has gained momentum Holler and Levinson [6]discussed the intricacies of multimodal language processing in human communication, and the fusion of information across modalities, known as multimodal fusion, augments emotional insights, refining the accuracy of outcomes [7]. Dobrišek et al. (2013) In recent years, Multimodal Sentiment Analysis (MSA) has garnered increasing attention, with researchers like Zadeh et al., & Tsai et al. [8-9] leading the discourse. This approach leverages multiple data modalities, demonstrating greater resilience and improved results, particularly when navigating the complexities of social media content. As online user-generated content flourishes, MSA finds applications expanding into areas like risk management, video comprehension, and transcription. However, MSA is not without its challenges, as identified by A. Zadeh et al[10], Incorporating diverse linguistic information, we integrate Incorporate acoustic and linguistic indicators into the model. Moreover, we employ inter-modality constructive learning to produce discriminative

1PhD Research scholar Department of Computer Science & Engineering, Integral University, Lucknow, India
2 Associate Professor Department of Computer Science & Engineering, Integral University, Lucknow, India
E-mail: 1mdusmankhhan@gmail.com, 2faiyaz.ahamad@yahoo.com

multimodal representations and emerges as a cornerstone, with recent contributions, such as insights from Hazarika [4], emphasizing that Unified-modal representations should embody both consistent and complementary data facets.

Our exploration categorizes existing methodologies into two paradigms: forward guidance and backward guidance. Forward guidance delves into crafting interactive modules for holistic cross-modal information capture, often grappling with nuanced capturing of modality-specific details. In contrast, backward guidance methods, as proposed by researchers like Lin, M and Hazarika [2,4], integrate additional loss functions facilitating representations that seamlessly merge both consistent and Unique modal characteristics. However, such methods often necessitate intricate weight balancing in their overarching loss functions, heavily contingent on human expertise.

The most significant contributions of this Study Include:

i. **Introduction of Unified-MSE Framework:** We introduce a groundbreaking multimodal sentiment knowledge-sharing framework, Unified-MSE, which harmonizes Sentiment Analysis across Modalities (MSA) and Conversational Emotion Detection (ERC)tasks. This innovative method capitalizes on the intrinsic similarities and complementarities between sentiments and emotions, enhancing predictive capabilities.

ii. **Integration of Multimodal Representation:** Our approach involves fusing multimodal representation by incorporating acoustic and visual signals into the model to integrate multi-level textual information. Additionally, we use inter-modality contrastive learning to produce discriminative multimodal representations, facilitating a more nuanced understanding of sentiment and emotion.

iii. **State-of-the-Art Performance:** The experimental results highlight the effectiveness of Unified-MSE, benchmark has been established across four widely recognized public the datasets, including MOSI, MOSEI, MELD, and IEMOCAP, are utilized for both MSA and ERC tasks. This achievement highlights to show effectiveness and flexibility of proposed framework.

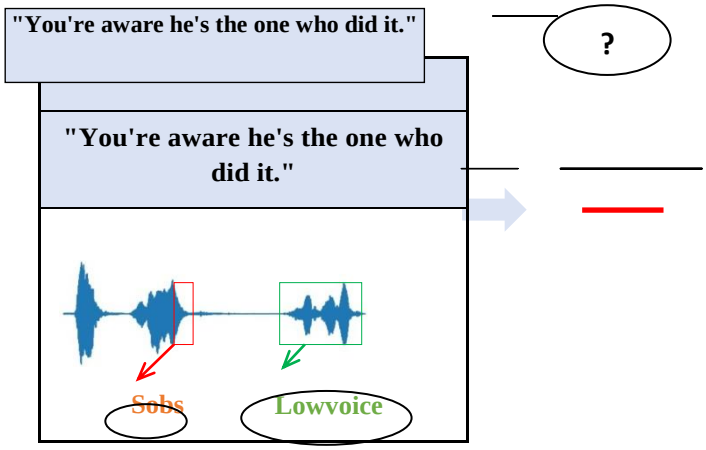


Fig 1: Instance to demonstrate the Inter modal approach interaction among textual & acoustic modalities.

2. Related Work

Multimodal Sentiment Analysis (MSA)

Multimodal Sentiment Analysis (MSA) offers a nuanced approach to sentiment analysis, aiming to not only determine sentiment polarity but also to measure the intensity of that sentiment. Introduced by Morency et al. [10], MSA harnesses multiple data modes, such as text, audio, and visual cues, to provide a richer sentiment understanding.

1. Multimodal Fusion: Historically, multimodal fusion centered on geometric manipulations within feature spaces [8]. This involved integrating features from various modalities in a geometrically consistent manner. However, recent advancements have ushered in more sophisticated techniques. Hazarika [4] introduced the reconstruction loss method, refining fused modality representations. Similarly, advanced the field with hierarchical mutual information maximization, spotlighting the most informative aspects from each modality.

2. Modal Consistency & Translation: Ensuring consistency across modalities is paramount in multimodal datasets. pioneered a multi-task joint learning approach, ensuring consistent sentiment representations across data types[9]. In contrast, delved

into techniques translating sentiment data between modalities, aiming for a harmonized sentiment perspective.

3. Multimodal Alignment: The alignment of sentiments across different data types remains crucial. leveraged cross-modality representations to align sentiment cues [11] elevated this with multi-scale modality representation, facilitating nuanced alignment across diverse data granularity.

4. Multimodal Context: Context is pivotal in sentiment analysis pioneered context-aware consideration mechanisms, allowing models to hone in on pertinent contextual data. MeanwhileM. U. Khan and F. Ahamad,[12] introducea multi-modal attention model, as described by Poria et al [8], employs a recurrent architecture featuring multi-level attentions, effectively capturing subtle contextual nuances in meaning.

3. Conversational Emotion Recognition

Emotion Recognition in Conversations (ERC) explores thoroughly into the task of perceptive emotions within conversational data. As conversations are dynamic, recognizing emotions in this realm presents unique challenges.

1. Multimodal Fusion: With the rise of multimodal machine learning, ERC has thrived. Notably connected graph of neural networks, modeling intricate dependencies between utterances and speakers, enhancing emotion recognition depth. [13]

2. Context Integration: Emotion recognition's efficacy in conversations hinges on context. employed graph structures to encapsulate the conversational context [5,14]. Furthermore, Mao et al [15] introduced affective changes., modeling the temporal shifts of emotions within conversations.

3. External Knowledge: Incorporating external knowledge augments ERC's capabilities. tapped into transfer learning [4]. Concurrently, S. Ghosh et. al [16] integrated commonsense knowledge, enriching ERC's comprehension of human emotions.

3.1 Unified-Framework

The convergence of disparate tasks into Unified-frameworks signifies modern machine learning's evolution. T5, championed by Raffel et al [17], epitomizes this trend, offering sentences on various NLP tasks. Harnessing this momentum, our research endeavors to integrate MSA and ERC within T5, aspiring to craft a embedding space, refining sentiment and emotion comprehension.

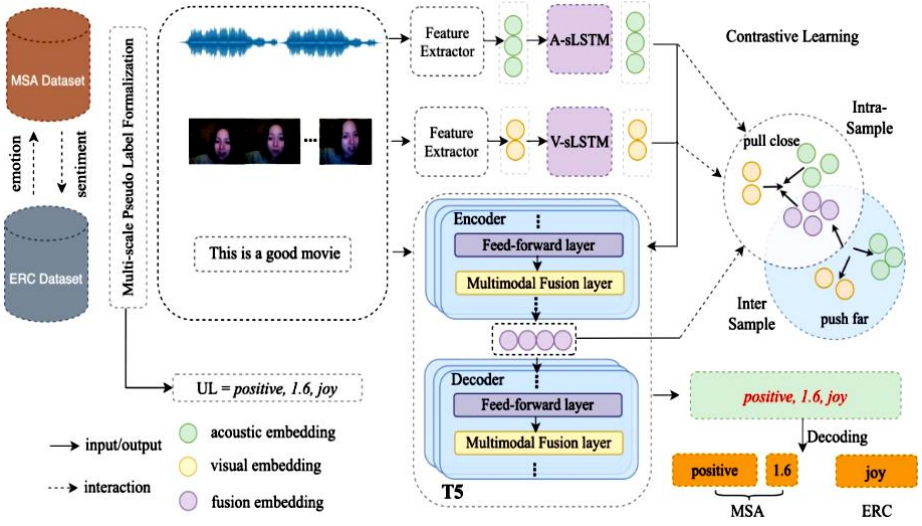


Fig 2: The overview of Unified-MSE

4. Methodology

4.1 Comprehensive Structure of Unified-MSA

In the intricate design of Unified-MSA, depicted in Figure 2, the architecture unfolds through a series of critical phases like formalizing the task, fusing pre-trained modalities, and implementing inter-modality contrastive learning. Initially, we engage in the offline processing of labels for MSA and ERC tasks, resultant them into a universal label (UL) format. Subsequently, we explore into the extraction of audio and text features, employing unified feature extractors that span across data-sets. Once these features are at our disposal, we navigate through the realms of long-term contextual insight by channeling them individually through dedicated LSTM pathways.

For the textual modality, the renowned T5 takes the stage as the encoder, diligently absorbing the nuanced contextual information encapsulated within the sequences. A distinctive facet of our approach is the seamless integration of multimodal fusion layers into the T5 architecture. These fusion layers are strategically embedded following the feed-forward layer within multiple Transformer layers of the T5 model.

Moreover, our innovation extends to the incorporation of inter-modal contrastive learning. This intricate process plays a crucial role in discerning and differentiating the multimodal fusion representations across various samples. The essence of contrastive learning lies in its mission to minimize the gap between modalities within the same sample, fostering cohesion, while simultaneously pushing the representations of different samples further apart. It's a delicate dance that refines the nuanced relationships between modalities and samples, ultimately enhancing the overall efficacy of Unified-MSA.

4.2 Task Formalization:

Involves processing multimodal signals, denoted as $I_i = \{I_i^t, I_i^a\}$where $I_i^m, m \in \{t, a\}$, represents unimodal raw sequences from video fragment i. Here, $\{t, a\}$refer to the three modalities: linguistic, acoustic, for multimodal sentiment analysis (MSA), the objective is to forecast the actual number $y_i^r \in R$ (The formula $y_i^r \in R$ represents that $y_i^r$is a real number, where R denotes the set of all real numbers), reflecting sentiment strength. Simultaneously, for emotion recognition in conversations (ERC), the objective is to forecast the emotional class of respective utterance and both MSA and ERC share similarities in the input characteristics, model structure, and label space due to task validation.

In this setup, we handle the input of dialogue text and modal features, while also making sure that the labels for both MSA and ERC tasks are the same by turning them into universal labels. This integrated approach streamlines the tasks, creating a cohesive framework that allows for joint consideration of sentiment strength prediction and emotion category identification. Furthermore, MSA and ERC are treated as a cohesive task, providing a unified and consistent foundation for modeling and analysis.

4.3 Input Formalization

Understanding the nuances of human emotions and intentions in conversations necessitates a deep comprehension of contextual information [17,18]. To capture this essence, we adopt a structured approach:

1. **Concatenation of Utterances**: We believe that the richness of context can be best grasped by amalgamating the current utterance u_i, with its adjacent 2-turn utterances, represented as $\{u_i - 1, u_i - 2\}$, and subsequent 2-turn utterances, denoted as $\{u_i + 1, u_i + 2\}$. This collection is presented as a raw text sequence: $I_t^i = [u_i - 2, u_i - 1, u_i + 2]$
2. **Segment Identification**: To distinguish the central utterance u_i, from its surrounding context, we introduce a segment identifier, S_i^t The formulation is represented as:

   $$S_i^t = [\underbrace{0, \dots, 0}_{}, \underbrace{1, \dots, 1}_{}, \underbrace{0, \dots, 0}_{}] ----(1)$$

   $\{z\}u_i - 2, u_i - 1\{z\}u_i\{z\}u_i + 1, u_i + 2$ Here, the $\{z\}$notation signifies a separation between utterances.
3. **Textual Modality Processing**: The aforementioned structured utterances undergo further processing to align with the textual modality, represented as I_i^t.
4. **Acoustic Feature Extraction**: Utilizing the librosa library, we transform raw acoustic signals into numerical sequential vectors. This process facilitates the extraction of Mel-spectrograms, a foundational tool in contemporary audio analysis, offering insights into the short-term power spectrum of sound.

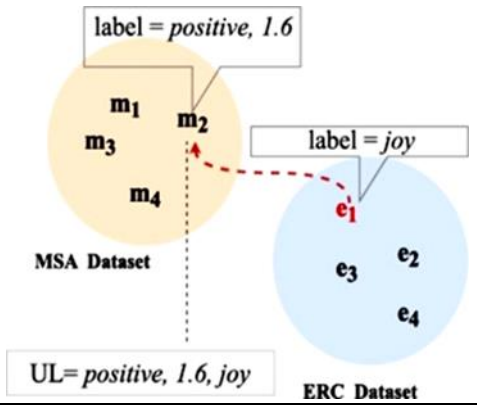


Figure 3: The technique use in creation of a universal label (UL) is illustrated, where the line marked with red dashes indicating that e_1 is the illustration most closely related in terms of semantic similarity to m_2.

4.4 Experimental Configuration

4.4.1 Datasets

We embarked on experiment leveraging four renowned benchmark datasets encompassing Multimodal Sentiment Analysis (MSA) and Conversation-based Emotion Recognition (ERC). The datasets include Rewrite: The MOSI dataset (Multimodal Opinion-level Sentiment Intensity) created by Zadeh et al [18], the MOSEI dataset (Multimodal Opinion Sentiment and Emotion Intensity) introduced by Zadeh et al [10], the MELD dataset (Multimodal Emotion Lines Data-set) developed by Poria et al. [19], and the IEMOCAP database (Interactive Emotional Dyadic Motion Capture) established by Busso et al [20].MOSI comprises 2,199 video segments, each meticulously Annotated with sentiment scores ranging from -3 to +3, indicating both the sentiment polarity and its intensity, MOSEI stands as an upgraded version of MOSI., boasts 22,856 movie review snippets sourced from YouTube. Although MOSEI offers both sentiment and emotion annotations, for this study, we solely rely on its sentiment annotations. Importantly, there's a clear demarcation between MOSI and MOSEI, distinct data collection and labeling methodologies were employed for each dataset.

Table 1: Provides a concise overview of the MOSI, MOSEI, MELD, and IEMOCAP datasets, offering insights into data distribution and label composition. 'Senti.' signifies sentiment polarity, while 'Emo.' represents emotion categories.

Dataset	Training Samples	Validation Samples	Test Samples	Sentiment Polarity Labels	Emotion Category Labels
MOSI	1284	229	686	Present	Not Present
MOSEI	16326	1871	4659	Present	Not Present
MELD	9986	1108	2610	Not Present	Present
IEMOCAP	5354	528	1650	Not Present	Present

IEMOCAP, on the other hand, is comprised of 7,532 samples. Drawing inspiration from prior research our focus for emotion recognition encompasses six distinct emotions: joy, sadness, anger, neutrality, excitement, and frustration. MELD houses 13,707 clips of multi-party dialogues, labeled in accordance with Ekman's six universally acknowledged emotions include joy, sadness, fear, anger, surprise, and disgust.

4.4.2 Acoustic characteristics and multimodal configuration:

In the research, we utilize each segment which is characterized by a 74-dimensional feature vector, which include 12 Mel-frequency cepstral coefficients (MFCCs), pitch and segmentation topographies, maxima dispersion quotients, glottal source and peak slope parameters. This is followed by obtaining alignment at the level of individual words. Features, we utilize P2FA Yuan and Liberman [19] conducted Identification of speakers within the SCOTUS corpus to discover the timestamps corresponding to each word. Subsequently, we calculate the mean of the audio features corresponding to the identified timestamps of each word. To maintain alignment consistency with the text mode, zero vectors are employed to pad the audio sequences.

5. Evaluation Benchmark

In the evaluation of MOSI and MOSEI datasets, we employed diverse metrics for a thorough model assessment. The Mean Absolute Error (MAE) gauged predictive accuracy, while Pearson Correlation (Corr) measured alignment with actual data trends. The Seven-Class Classification Accuracy (ACC-7) provided insights into multi-class performance, whereas Binary Classification Accuracy (ACC-2) focused on binary distinctions. The F1 Score balanced precision and recall across various classifications, for the MELD & IEMOCAP datasets, our focus primarily revolved around Accuracy (ACC) and Weighted F1 (WF1) metrics to ensure fairness in evaluations.

6. Results

In our comparative analysis across datasets including MOSI, MOSEI, IEMOCAP, and MELD, Unified MSE emerges as a frontrunner, surpassing the current state-of-the-art (SOTA) benchmarks. Specifically, Unified MSE demonstrates notable enhancements in various metrics. For instance, the ACC-2 metrics witness a boost of 1.65% for MOSI and 1.16% for MOSEI, while the ACC metrics show improvements of 2.6% for MELD and 2.35% for IEMOCAP. Additionally, there's a

commendable rise in F1 scores, with MOSI and MOSEI benefiting by 1.73% and 1.29%, respectively. It's worth noting that while early research endeavors like LMF and TFN provided comprehensive coverage across all datasets, recent methodologies have often been limited to specific datasets or particular metrics. In contrast, Unified MSE offers a holistic approach, addressing both MSA and ERC tasks across the board, thereby underscoring its prowess as a unique and superior framework within the realm of sentiment analysis & emotion recognition.

I. ABLATION STUDY ON UNIFIED-MSE

Firstly, by progressively eliminating individual or multiple modalities from the multimodal signals, we assessed their impact on the model's performance. Notably, excluding either the visual or acoustic modalities, or both, resulted in a noticeable decline in performance metrics. This underscores the significance of non-verbal cues, such as visual and acoustic signals, in MSA tasks, highlighting the synergistic relationship between text, acoustic, and visual data. Intriguingly, of the two, the acoustic modality emerged as more pivotal for Unified MSE. Subsequently, we examined the role of specific components within Unified-MSE, namely the PMF and CL modules. Their exclusion led to adverse effects, manifested as increased MAE and diminished Corr scores, underscoring their crucial role in facilitating effective learning of multimodal representations.

Table 2: Analysis of Unified-MSE's Ablation on MOSI: In this study, we examined the impact of removing specific modalities from Unified-MSE on the MOSI dataset.

Modality	MAE	Corr	ACC-2	F1-Score
Unified-MSE	**0.702**	**0.819**	**86.05/86.44**	**85.92/86.37**
Acoustic (A)	0.743	0.797	84.07/85.66	84.11/85.60
Visual (V)	0.736	0.805	84.59/85.61	84.92/85.75
Acoustic (A), Visual (V)	0.729	0.785	83.98/85.34	83.79/85.26
P.M.F	0.744	0.790	85.26/86.37	85.13/86.18
C.L	0.727	0.808	85.46/86.69	85.47/86.63
IEMOCAP	0.713	0.795	85.28/86.59	85.27/86.55
MELD	0.722	0.776	84.05/84.96	84.50/84.64
MOSEI	0.775	0.727	80.68/81.22	81.35/81.83

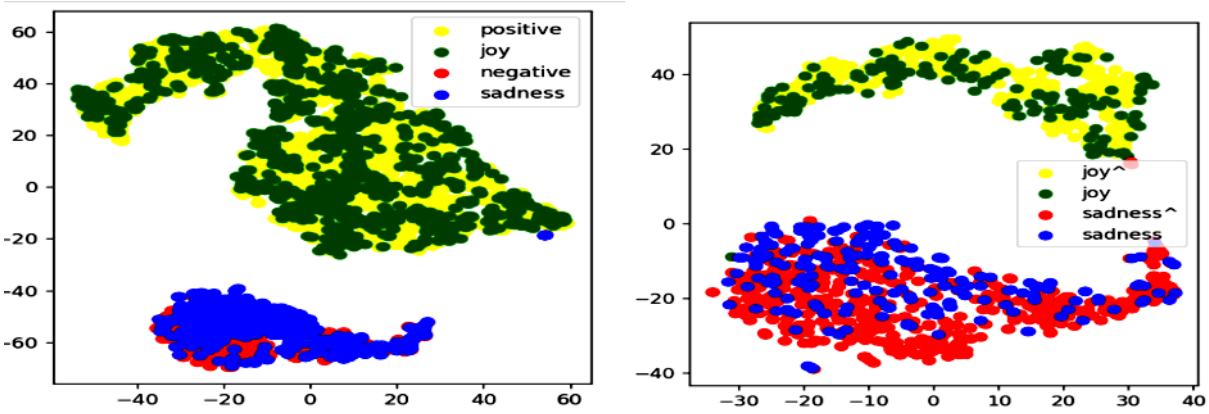
Furthermore, we performed experiments to assess the dataset's impact on Unified MSE's performance. Specifically, omitting individual datasets like --Using the training sets of IEMOCAP, MELD, and MOSEI to train the model, we evaluate its effectiveness on the MOSI test set discovered interesting insights. While excluding IEMOCAP and MELD resulted in a notable decline in performance metrics, particularly in MAE and Corr., suggesting that these datasets provide valuable information essential for the MSA task. Conversely, excluding MOSEI resulted in a decline across all metrics.

In our extensive evaluation on the MOSI dataset, we performed a sequence of ablation experiments to discern the importance of different modalities and components within Unified-MSE. The findings are presented in Table 2. Here, V represents the visual modality, while A represents the acoustic modality. Furthermore, PMF stands for pre-trained modality fusion, and CL corresponds to contrastive learning. Proposed Unified-MSE framework stands distinct from existing methodologies. Its efficacy and versatility are evidenced by the consistent improvements observed across various ablation scenarios, underscoring its potential applicability across diverse tasks.

7. Visualization

To assess the influence of Unified-MSE's Universal Label (UL and cross-task learning, we illustrate the multimodal representation F(j) derived from the final Transformer layer. For this purpose, we specifically select samples demonstrating Sentiment polarity (positive/negative) from the test set of MOSI, uttering sadness/happiness emotions from MELD's test set. The

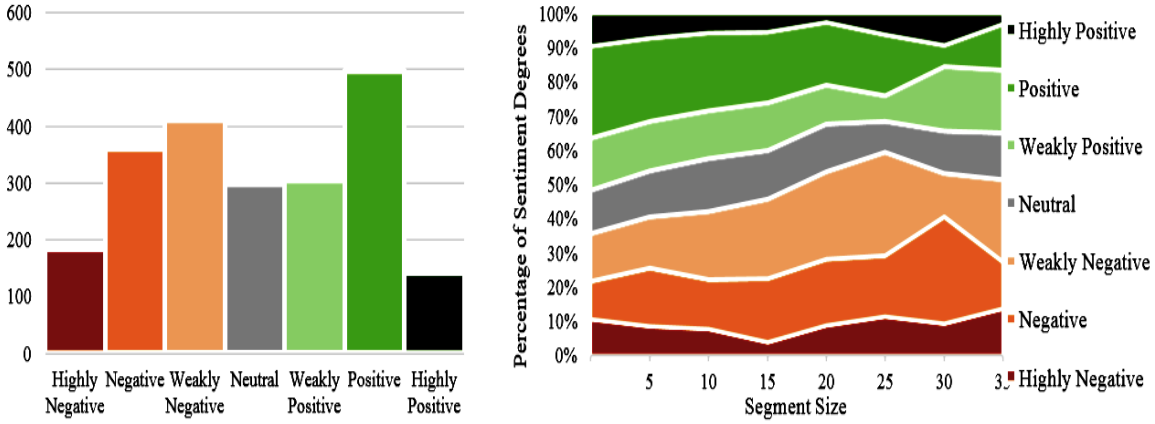
visual depiction of these representations is presented in Figure 4(a).



(a) (b)

Fig 4: The visualization compares multimodal fusion representations: (a) samples categorized by sentiment & emotion, where (b) samples displaying novel and synthesized emotions concurrently, with happiness and sadness representing the made emotions.

Furthermore, we also include MOSI samples that have generated emotions. The visualization compares joy/sadness samples with MELD instances bearing novel emotion labels of happiness/sadness in the embedding space in Figure 4(b). Notably, joy-labeled samples, whether original or synthesized based on the Universal Label (UL) share in a feature space. These outcomes highlight Unified-MSE's proficiency in learning how to represent data across various instances and underscore the complementary correlation between sentiment and emotion. Figure 5 depicts a histogram: 5(a) showcases sentiment distribution across the dataset, while 5(b) illustrates the proportion of each sentiment category relative to segment size, measured by the number of words in each opinion segment.



(a) **(b)**

Fig 5: Illustrates the sentiment distribution across the entire dataset in the left histogram. On the right, the graph displays the percentage of each sentiment degree per segment size, where the segment size represents the number of words in an opinion segment.

8. Discussion

In this study, the Unified-MSE framework is introduced and evaluated for sentiment analysis and emotion recognition tasks using diverse datasets. The COVAREP toolkit is employed to extract audio features, and P2FA aids in word-level alignment, ensuring consistency with zero vector padding. Evaluation metrics, including MAE, Pearson Correlation, ACC-7, ACC-2, and F1 Score, provide a comprehensive assessment across datasets. Unified-MSE demonstrates superior performance, surpassing existing benchmarks and showcasing a holistic approach to sentiment assessment and emotion identification. The ablation study emphasizes the crucial role within linguistic and acoustic modalities. The acoustic modality being particularly pivotal. Experiments excluding specific datasets underscore the significance of IEMOCAP and MELD for effective task performance.

Detailed insights from Table 2 highlight the impact of modality and dataset removal on Unified-MSE's performance, consistently affirming its superiority and versatility. The visualization section further validates the model's proficiency through visual representations of shared feature spaces among samples with similar emotions. So, Unified-MSE emerges as a robust and versatile framework, excelling in sentiment analysis and emotion recognition tasks. The study's meticulous evaluations and ablation studies underscore the framework's efficacy, especially in leveraging multimodal data for enhanced representation learning.

9.1 Theoretical Implications:

i. **Advancement in Multimodal Analysis:** The introduction of the Unified-MSE framework underscores a significant advancement in the integration of multimodal data for sentiment analysis & emotion recognition. This study bridges the gap between textual, acoustic, and visual modalities, offering a more holistic perspective on human communication.

ii. **Role of Acoustic Modality:** The ablation study's emphasis on the pivotal role of the acoustic modality highlights the significance of non-verbal cues in sentiment and emotion recognition. This finding enriches our theoretical understanding of the interplay between linguistic and paralinguistic elements in communication.

iii. **Dataset Significance:** The study's insights into the impact of specific datasets, such as IEMOCAP and MELD, on task performance contribute to a nuanced understanding of dataset selection and its implications for model generalization and effectiveness.

9.2 Practical Implications:

i. **Enhanced Model Performance:** The superior performance of the Unified-MSE framework, as demonstrated across diverse datasets and evaluation metrics, suggests its potential for practical applications. Organizations and researchers can leverage this framework to develop more accurate and reliable sentiment analysis and emotion recognition tools.

ii. **Optimized Data Utilization:** The study's emphasis on the importance of multimodal data and specific datasets informs practitioners about the optimal utilization of data sources. This knowledge can guide data collection and preprocessing efforts, ensuring that relevant modalities and datasets are prioritized.

iii. **Visualization for Interpretability:**The visual representations generated by the Unified-MSE framework offer practical tools for model interpretability and validation. Stakeholders can utilize these visualizations to gain insights into the model's decision-making process and to assess its alignment with human perceptions of sentiment and emotion.

9. Conclusion

This study introduces a psychological lens, emphasizing the feasibility and rationale behind jointly modeling sentiment and emotion. We unveil the Unified-MSE framework, a unified multimodal knowledge-sharing approach tailored for MSA and ERC tasks. Beyond merely capturing sentiment and emotion knowledge, Unified-MSE effectively aligns input features with output labels. we proposed an integrated method we conducted thorough experiments on four benchmark datasets, integrating multi-level textual features with acoustic and visual data representations, and employing inter-modality contrastive learning. We demonstrate SOTA (state of the art) outcomes across all evaluated indicators. Additionally, our visualizations of multimodal representations validate the significance of sentiment & emotion in the embedding space is explored in this research. We anticipate that the study will introduce a novel experimental paradigm, offering a fresh perspective to both the MSA & ERC research.

However, like any pioneering endeavor, our research is not without its constraints. A noteworthy point of consideration is our current reliance on the MELD and IEMOCAP datasets, leaving room for enriched insights from datasets like MOSI and MOSEI. Moreover, while our textual-centric approach has shown promise, the realm of acoustic modalities beckons exploration. These identified gaps not only underscore the nascent stage of our research but also serve as guiding stars for future endeavors. Looking ahead, our journey with the Unified-MSE framework is far from over. The roadmap ahead is illuminated with opportunities to broaden our dataset horizons, refine computational methodologies, and embrace interdisciplinary collaborations. By combining the technical with the human-centric insights from psychology and linguistics, we aspire to craft models that resonate more authentically with the essence of human emotion and sentiment. Additionally, as we tread this path, ethical considerations and human-centric evaluations will remain at the forefront, ensuring that our advancements uphold the sanctity and sensitivity of human communication. Through this holistic approach, we envision not just advancements in computational models but a deeper, more resonant understanding of the human narrative.

10. Acknowledgement

In accordance with the university doctoral studies and research guidelines, we recognize the allotment of **Manuscript Communication Number [IU/R&D/2024-MCN0002366]** to this article. This unique identifier aids in facilitating communication and monitoring of our research during the publication process. We extend our gratitude to all who have contributed to this project.

References

[1] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L. P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 163–171.

[2] M. Lin et al., "Modern dialogue system architectures,"*Journal of Conversational AI*, vol. 8, no. 2, pp. 45-60, 2020.

[3] K. Lin and J. Xu, "Emotion recognition in conversational agents,"*Dialogue Systems Journal*, vol. 14, no. 1, pp. 15-29, 2019.

[4] N. Majumder et al., "Multimodal sentiment analysis using hierarchical fusion with context modeling,"*Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.

[5] T. Ahmad, S. U. Ahmed, and N. Ahmad, "Detection of Depression Signals from Social Media Data," in *Smart Connected World: Technologies and Applications Shaping the Future*, 2021, pp. 191-209.

[6] J. Holler and S. C. Levinson, "Multimodal language processing in human communication,"*Trends in Cognitive Sciences*, 2019.

[7] S. Dobrišek et al., "Towards efficient multi-modal emotion recognition,"*International Journal of Advanced Robotic Systems*, vol. 10, no. 1, p. 53, 2013.

[8] B. Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.

[9] Y. Tsai et al., "Cross-modality representation in sentiment analysis,"*Multimodal Systems Journal*, vol. 16, no. 3, pp. 40-54, 2019.

[10] Zadeh et al., "multi-attention recurrent network for human communication comprehension," in Thirty-Second AAAI Conference on Artificial Intelligence.

[11] R. Li et al., "Towards discriminative representation learning for speech emotion recognition," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[12] M. U. Khan and F. Ahamad, "An Affective Framework for Multimodal Sentiment Analysis to Navigate Emotional Terrains,"*Telematique*, vol. 23, no. 01, pp. 70-83, 2024.

[13] Joshi et al., "Inter/intra dependencies modeling in dialogue systems,"Journal of Multimodal Systems, vol. 13, no. 1, pp. 12-28, 2022.

[14] Li et al., "Contextual graph structures for emotion modeling,"Journal of Multimodal Systems, vol. 14, no. 3, pp. 56-71, 2021.

[15] X. Tan, M. Zhuang, X. Lu, and T. Mao, "An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model," in *IEEE Access*, vol. 9, pp. 15860-15871, 2021, doi: 10.1109/ACCESS.2021.3052566.

[16] S. Ghosh et al., "Context and Knowledge Enriched Transformer Framework for Emotion Recognition in Conversations," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533452.

[17] C. Raffel et al., "T5: A unified framework for NLP tasks,"*Journal of Natural Language Processing*, vol. 26, no. 4, pp. 1302-1317, 2020.

[18] Zadeh et al., "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos,"IEEE Intelligent Systems, vol. 31, no. 6, pp. 82-88, 2016, doi: 10.48550/arXiv.1606.06259.

[19] S. Poria et al., "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 527–536.

[20] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database,"*Language resources and evaluation*, vol. 42, pp. 335-359, 2008.