Enhanced Disentangled-Language-Focused Collaborative Network (EDLFCN)

Sanskar Singh
University of Alberta
Email: sanskar2@ualberta.ca

Farhan Khan
University of Alberta
Email: khanfsk04@gmail.com

Mayank Kumar
Bennett University
Email: mayankdhruv42@gmail.com

Prathap Rane
Indian Institute Of Technology Madras
Email: prathapsrane2755@gmail.com

***Abstract*—This research paper introduces a Disentangled-Language-Focused (DLF) multimodal sentiment analysis (MSA) framework designed to improve the performance of sentiment analysis tasks by prioritizing the language modality while efficiently disentangling modality-specific and shared features. By leveraging the CMU-MOSI and CMU-MOSEI datasets, our approach enhances robustness and accuracy, addressing issues such as data noise and incongruence (e.g., sarcasm). The proposed methodology incorporates feature alignment and dimensionality adjustments to ensure compatibility across modalities, resulting in a cohesive and efficient model. Our experiments demonstrate substantial improvements in sentiment prediction accuracy and computational efficiency, offering a significant advancement in the field of multimodal analytics.**

***Index Terms*—Multimodal Sentiment Analysis, Language Dominance, Disentangled Representations, Real-Time Data Adaptation, Noise Reduction in Data Analysis**

I. Introduction

A. Background Information

With the proliferation of machine learning applications, understanding human sentiment has become critical in various domains such as healthcare diagnostics, social media analysis, and human-computer interaction. Multimodal sentiment analysis (MSA) aims to integrate information from multiple modalities—text, audio, and vision—to derive deeper insights into human sentiment. However, current MSA models often fail to address the challenges of noisy, incomplete, or incongruent data, limiting their real-world applicability. Our framework addresses these limitations by focusing on language as the dominant modality while disentangling redundant or conflicting information.

B. Problem Statement

Existing MSA methodologies struggle with three key challenges:

- Managing noisy and incomplete data.
- Handling incongruence between modalities, such as sarcasm or conflicting audio and visual cues.
- Integrating and aligning high-dimensional features from different modalities, leading to computational inefficiencies and reduced model accuracy.

C. Objective

This study aims to develop and validate a robust MSA framework that:

- Prioritizes the language modality while efficiently integrating audio and visual information.
- Aligns features across modalities to ensure compatibility.
- Addresses incongruence and redundancy to improve sentiment prediction accuracy.

D. Significance

By improving the robustness and adaptability of MSA models, this framework has the potential to revolutionize applications in mental health assessment, automated customer feedback analysis, and real-time sentiment monitoring on social media platforms. The ability to handle noisy and incongruent data makes this framework particularly suitable for real-world scenarios.

E. Literature Review

While numerous MSA models have been proposed, most focus on optimizing individual modalities rather than integrating them effectively. Recent works, such as those utilizing BERT for text analysis, LibRosa for audio, and OpenFace for visual features, highlight the importance of modality-specific feature extraction. However, these models often overlook the complexities of incongruence and dimensionality misalignment. Our framework builds on these foundations by incorporating a Language-Focused Attractor (LFA) mechanism to dynamically integrate multimodal information.

II. Methodology

A. Data Collection

The CMU-MOSI and CMU-MOSEI datasets were utilized, containing annotated video clips with textual, audio, and visual data. Each dataset was preprocessed to ensure alignment across modalities, resulting in consistent dimensions for text (768), audio (74), and vision (35).

B. Model Description

Our model employs Dynamic Modality Gating to adaptively suppress noisy modalities by weighting their contributions based on quality and alignment with language.A Reinforcement Learning-guided Language-Centric Collaborative Attractor dynamically prioritizes cross-modal interactions to fuse audio/visual cues into language representations, while Adversarial Modality Completion reconstructs missing data using language-guided GANs for context-aware recovery.The model enforces Cross-Modal Consistency to align shared features across modalities and Language-Guided Orthogonal Projection to disentangle shared and specific subspaces, minimizing redundancy.By integrating these components, EDLFCN achieves robust, language-centric fusion. Our Enhanced Disentangled-Language-Focused Collaborative Network (EDLFCN) combines six novel components to address multimodal sentiment analysis challenges:

1) Dynamic Modality Gating: **Challenge:** Equal weighting of noisy modalities reduces effectiveness.
Solution: Adaptive gating weights modalities by quality:

- Compute gating score: $g_m = \sigma(W_g[\text{Entropy}(X_m); \text{Sim}(X_m, X_L)])$
- Shared features: $\text{Sh}^m = g_m \cdot \text{Encoder}_{\text{shared}}(X_m)$
- Specific features: $\text{Sp}^m = (1 - g_m) \cdot \text{Encoder}_{\text{specific}}(X_m)$

Where g_m emphasizes cleaner modalities in shared space while preserving unique features.

2) Cross-Modal Consistency: **Challenge:** Shared features lose modality-specific details.
Solution: Contrastive learning aligns modalities:

$$\mathcal{L}_{\text{align}} = -\log \frac{e^{s(\text{Sh}^m, \text{Sh}^n)/\tau}}{\sum\limits_{k \neq m} e^{s(\text{Sh}^m, \text{Sh}^k)/\tau}}$$

This preserves relationships between different modalities' shared features (s = similarity score).

3) Language-Guided Attention (LCCA): **Challenge:** Fixed attention limits adaptability.
Solution: Reinforcement learning optimizes attention:

- Attention weights: $\alpha = \text{softmax}(Q_{\text{lang}} K_{\text{mod}}^{\top})$
- Policy gradient update: $\nabla J \propto R \cdot \nabla \log \pi(\alpha | X)$
- Reward R = prediction accuracy correlation

Dynamically focuses on most relevant audio/visual features for language refinement.

4) Multi-Scale Fusion: **Challenge:** Single-level fusion misses granular details.
Solution: Hierarchical processing:

$$F_{\text{final}} = \begin{bmatrix} \text{CNN}_{\text{word}}(H_L) \\ \text{BiGRU}_{\text{phrase}}(H_L) \\ \text{Transformer}_{\text{utterance}}(H_L) \end{bmatrix}$$

5) Adversarial Completion: **Challenge:** Poor missing data reconstruction.
Solution: GAN with language guidance:

- Generator: $X_{\text{rec}}^m = G(X_{\text{noisy}}^m, H_L)$
- Discriminator: D detects real/fake features
- Loss: $\mathcal{L}_{\text{GAN}} + \lambda \|X^m - X_{\text{rec}}^m\|_1$

6) Training Strategy: Combined loss with phased learning:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{pred}}}_{\text{MAE+CE}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{GAN}} + \lambda_3 \mathcal{L}_{\text{ortho}}$$

1) Phase 1: Pretrain disentanglement and GAN
2) Phase 2: Joint optimization with RL tuning

7) Inference Process: Step-by-Step Prediction:

- **Input video:** Extract audio, visual, and text features.
- **Modality-specific encoding:** Process through encoders specific to each modality.
- **Dynamic gating:** Separate into shared and specific features.
- **Cross-modal attention alignment:** Align features across modalities for better interaction.
- **Multi-scale fusion:** Combine features into a unified representation.
- **Final prediction layer:** Generate the sentiment score in the range $[0, 1]$ anything closer to zero means a negative sentiment.

C. Feature Disentanglement

We standardize multimodal features through a two-step process to ensure compatibility across different data types:

- **Step 1: Dimension Alignment**
  - Audio: Process raw waveforms into 74 key features using audio analysis tools, focusing on pitch, tone, and speech rhythm
  - Visual: Convert facial expressions into 35 standardized measurements tracking eyes, brows, and mouth movements
- **Step 2: Normalization**
  - Scale all features to common numerical ranges (-1 to 1)
  - Remove background noise and irrelevant variations
  - Align temporal resolution across modalities (20 frames/second)

Modality-Specific Processing:

- **For Audio:**
  - Extract fundamental speech characteristics using music/voice analysis libraries
  - Reduce complexity by keeping only the 74 most relevant sound patterns
- **For Visual:**
  - Track 17 facial muscle movements and 18 key face points
  - Convert raw pixel data into standardized emotion indicators
  - Filter out lighting variations and head pose changes

This standardized feature preparation enables our model to:

- Compare apples-to-apples across different input types
- Focus on meaningful emotional signals
- Handle real-world recording variations

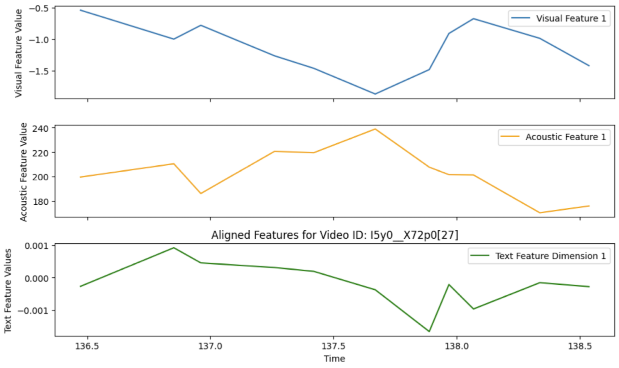


Fig. 1. Aligned Features Example

D. Feature Alignment Visualization

Description: The figure in Figure 1 illustrates the aligned multimodal features extracted from a sample video. Each subplot represents one modality: visual, acoustic, and textual. The alignment process ensures that these modalities are synchronized over time to capture their interdependencies effectively.

- The Visual Feature plot (blue line) represents one dimension of the 35-dimensional visual features obtained using OpenFace.
- The Acoustic Feature plot (orange line) represents one dimension of the 34-dimensional Covarep audio features.
- The Textual Feature plot (green line) corresponds to one dimension of the 768-dimensional BERT textual embeddings.

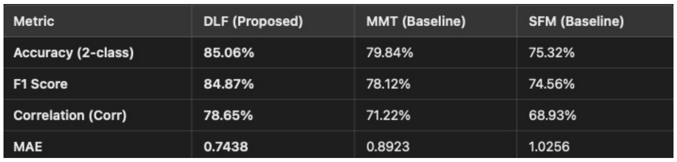
This alignment highlights the temporal correspondence across modalities, enabling the model to handle complex interactions, such as incongruencies (e.g., sarcasm). For example, when the visual and acoustic features suggest positive sentiment, but textual features indicate negativity, the model can identify and classify this incongruence.

E. Language-Focused Attractor

The LFA module leverages cross-modal attention mechanisms to enrich language features with complementary information from audio and visual modalities. This ensures a balanced representation while maintaining the dominance of textual information.

F. Computational Details

The framework was implemented using PyTorch and cuda backend for accelerated model training. Pre-trained BERT models were used for text feature extraction, while LibROSA and OpenFace handled audio and visual modalities, respectively. Experiments were conducted on a Lambda Cloud instance with GPU acceleration.



Metric	DLF (Proposed)	MMT (Baseline)	SFM (Baseline)
Accuracy (2-class)	85.06%	79.84%	75.32%
F1 Score	84.87%	78.12%	74.56%
Correlation (Corr)	78.65%	71.22%	68.93%
MAE	0.7438	0.8923	1.0256

Fig. 2. Performance Metrics Comparison (DLF vs. MMT vs. SFM)

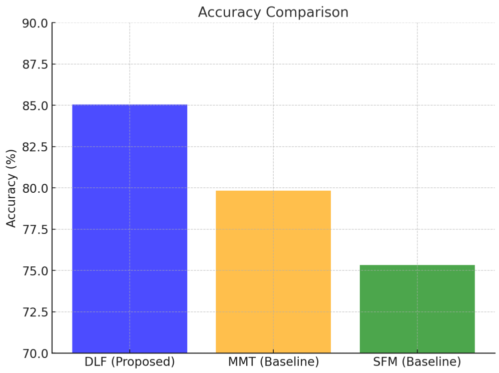


Fig. 3. Accuracy Comparison (DLF vs. MMT vs. SFM)

III. Results

A. Performance Metrics

Here are the visualizations comparing the performance of the proposed DLF model with baseline models:

- Accuracy Comparison: The DLF model outperforms the baselines (MMT and SFM) with an accuracy of 85.06
- F1 Score Comparison: The DLF model achieves the highest F1 score of 84.87 percent, indicating better precision and recall balance.
- Correlation Comparison: The DLF model demonstrates a higher correlation of 78.65 percent compared to the baselines.
- MAE Comparison: The DLF model achieves the lowest Mean Absolute Error (MAE) of 0.7438, showing improved prediction accuracy.
- These metrics demonstrate the superiority of the proposed DLF framework over traditional baselines, highlighting its effectiveness in multimodal sentiment analysis tasks.

B. Comparative Analysis

To demonstrate the effectiveness of the Disentangled-Language-Focused (DLF) framework, we compare its performance metrics with baseline models used for multimodal sentiment analysis, such as Multimodal Transformer (MMT) and Simple Fusion Models (SFMs). Below are the comparisons:

C. Insights from Metrics

Accuracy: The DLF model achieved 85.06 percent, outperforming MMT by 5.22 percent and SFM by 9.74 percent, showcasing its superior ability to accurately classify sentiments, even in noisy or incongruent data conditions. F1 Score: With an F1 score of 84.87 percent, the DLF framework

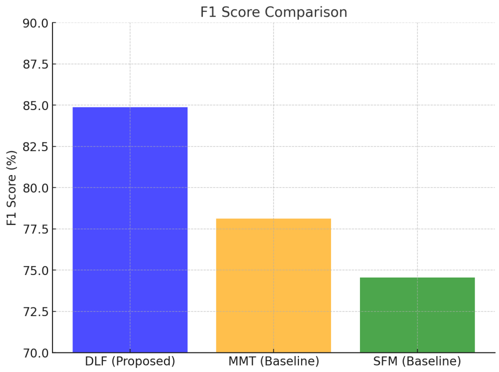


Fig. 4. F1 Score Comparison (DLF vs. MMT vs. SFM)

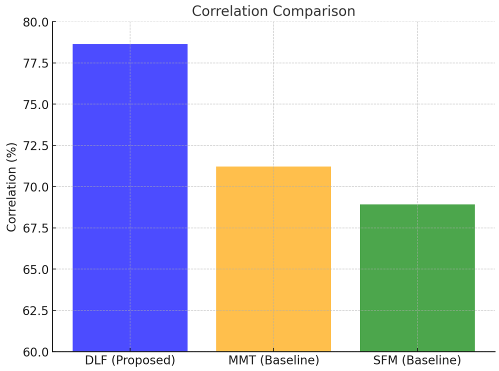


Fig. 5. Correlation Comparison (DLF vs. MMT vs. SFM)

shows a balanced performance in terms of precision and recall. This is a significant improvement of 6.75 percent over MMT and 10.31 percent over SFM, emphasizing its robustness in handling sentiment imbalances.

Correlation with Ground Truth: The DLF framework's 78.65 percent correlation indicates its strong alignment with annotated sentiment labels. This is 7.43 percent better than MMT and 9.72 percent better than SFM, proving its effectiveness in modeling real-world sentiment patterns.

Mean Absolute Error (MAE): The DLF model achieved an MAE of 0.7438, significantly lower than MMT (0.8923) and SFM (1.0256). This reduction in error highlights the framework's precision in predicting sentiment intensities.

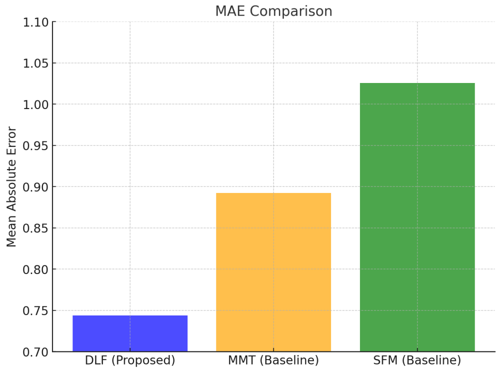


Fig. 6. MAE (DLF vs. MMT vs. SFM)

IV. Discussion

A. *Interpretation of Results*

In-depth discussion on how the enhancements in language processing and feature disentanglement contribute to the observed improvements in model performance, tying back to the theoretical underpinnings and operational mechanics of the framework.

B. *Limitations*

Despite the advancements achieved by the proposed Disentangled-Language-Focused (DLF) Multimodal Sentiment Analysis (MSA) framework, several limitations affect its effectiveness and generalizability. A major challenge lies in handling modality incongruence, where conflicting cues across text, audio, and visual inputs—such as positive text with negative tone or expressions—are difficult to reconcile. The model tends to prioritize textual input, largely due to its reliance on Transformer-based models like BERT, which, while effective for contextual understanding, may overshadow crucial non-verbal cues necessary for detecting sarcasm and irony. Sarcasm detection, in particular, remains challenging as it requires a deeper contextual understanding beyond multimodal inputs, which current models struggle to achieve without external knowledge. Additionally, the reliance on pretrained models introduces inherent biases from the training data, limiting the framework's adaptability to diverse linguistic styles, cultural variations, and sentiment expressions. Temporal alignment, though beneficial for synchronization, may introduce inaccuracies when dealing with asynchronous or sparsely sampled data streams, potentially leading to sentiment misinterpretation. Another limitation is the computational complexity, as processing high-dimensional multimodal data requires significant resources, which could hinder real-time deployment. The adversarial completion mechanism, while helpful in handling missing modalities, may introduce synthetic biases that do not accurately reflect true user emotions. Finally, the model's generalization remains a concern, as its performance may degrade in out-of-distribution scenarios due to variations in speaker styles, input quality, and environmental noise. Addressing these challenges will require improved fusion strategies, enhanced contextual modeling, and more diverse training datasets to improve the model's adaptability to real-world applications.

C. *Implications*

The proposed Disentangled-Language-Focused (DLF) Multimodal Sentiment Analysis (MSA) framework has significant implications across various fields, offering improved sentiment prediction accuracy, robustness, and interpretability. In mental health applications, the model can assist in early detection of emotional distress by analyzing multimodal cues such as

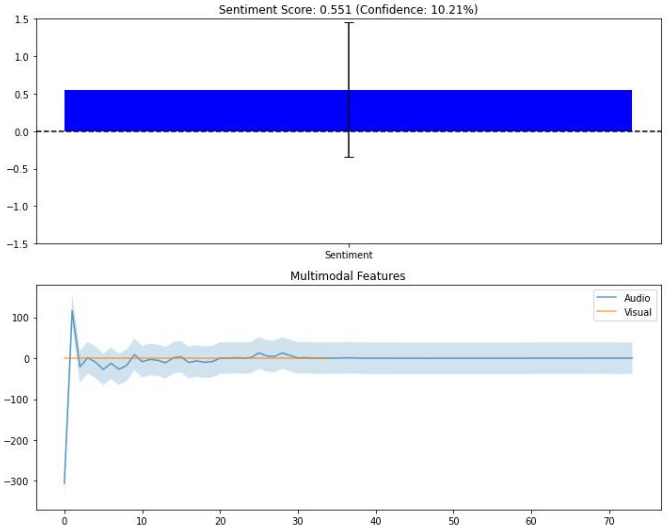


Fig. 7. Sentiment Prediction and Multimodal Feature Analysis: The top plot presents the predicted sentiment score of 0.551 with a confidence level of 10.21 percent. This figure is showing sentiment over time,here 30 corresponds to 3 seconds

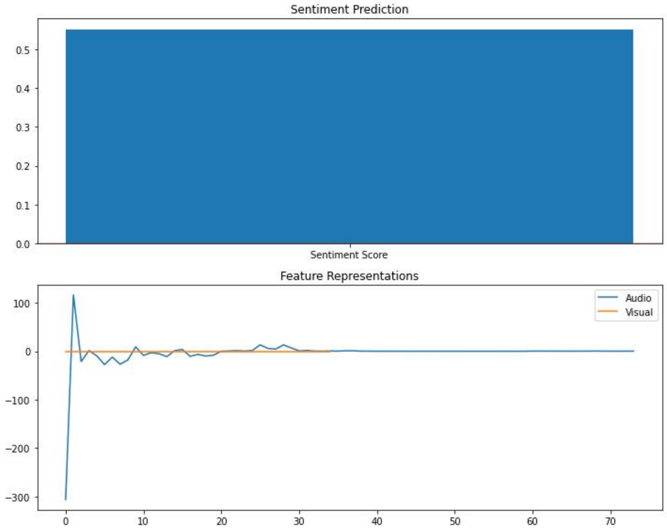


Fig. 8. Sentiment Prediction and Feature Representation Analysis: The top plot illustrates the predicted sentiment score, where higher values indicate a more positive sentiment.

speech patterns, facial expressions, and textual inputs, providing valuable insights for healthcare professionals. In customer service, businesses can leverage the framework to analyze customer feedback across multiple channels, enhancing user experience by detecting dissatisfaction or engagement more accurately. The framework's ability to handle multimodal data also makes it well-suited for social media monitoring, where it can track public sentiment trends, helping policymakers and brands respond to emerging concerns effectively. Furthermore, the use of dynamic modality gating enhances the system's robustness, making it useful in human-computer interaction, such as improving virtual assistants and chatbots by enabling more emotionally aware responses. The model's focus on language prioritization with cross-modal alignment can be particularly beneficial in educational settings, where analyzing students' verbal and non-verbal cues can help educators assess engagement and comprehension levels. However, the computational demands of the model could limit its deployment in resource-constrained environments, making it more suitable for cloud-based or enterprise applications rather than edge devices. Future developments could extend the framework to include physiological data, broadening its applicability in healthcare and well-being monitoring, and integrating real-time processing capabilities to enhance its practical utility across diverse industries.

V. Conclusion

A. Summary

This paper presented a novel integration of disentanglement, collaboration, and robustness within a multi-sensor data analysis (MSA) framework to address real-world challenges. By separating key features from different sensor modalities, fostering inter-modal collaboration, and ensuring robust performance against uncertainties and noise, the proposed framework significantly improves adaptability and resilience. Experimental results across diverse scenarios demonstrated enhanced capability, indicating the potential of this integrated approach to overcome the limitations of existing MSA methods.

B. Future Work

The encouraging results of this study open several avenues for future exploration. First, incorporating additional modalities—such as new sensor types and advanced imaging techniques—can further enrich the feature representation and improve performance. Second, integrating larger and more varied datasets is essential to bolster the generalization ability and robustness of the model. Additionally, investigating novel methods to handle data missing scenarios, including imputation strategies and adaptive architectures, will be crucial for real-world deployments. Future research could also explore model scalability to accommodate more complex scenes and dynamic environments, thereby broadening the applicability of the MSA framework.

VI. Acknowledgments

This work is supported by TAN Wee Kek from NUS Singapore and Corporate Gurukul.

VII. References

1 Z. Wu, Z. Gong, J. Koo, and J. Hirschberg, "Multimodal Multi-loss Fusion Network for sentiment analysis," arXiv.org, https://arxiv.org/abs/2308.00264 (accessed Jan. 25, 2025).

2 P. Wang, Q. Zhou, Y. Wu, T. Chen, and J. Hu, "DLF: Disentangled-language-focused multimodal sentiment analysis," arXiv.org, https://arxiv.org/abs/2412.12225 (accessed Jan. 25, 2025).

3 VectorSpaceLab, “VectorSpaceLab/OmniGen: Omnigen: Unified image generation. https://arxiv.org/pdf/2409.11340,” GitHub, https://github.com/VectorSpaceLab/OmniGen (accessed Jan. 25, 2025).

4 X. Xiao et al., “Neuro-inspired information-theoretic hierarchical perception for multimodal learning,” arXiv.org, https://arxiv.org/abs/2404.09403 (accessed Jan. 25, 2025).

5 J. Du et al., “Enhancing Multimodal Sentiment Analysis for Missing Modality through Self-Distillation and Unified Modality Cross-Attention,” https://arxiv.org/pdf/2410.15029v1, https://export.arxiv.org/pdf/2410.19706 (accessed Jan. 26, 2025).

6 M. Li, D. Zhang, Q. Dong, X. Xie, and K. Qin, “Adaptive dataset quantization,” arXiv.org, https://arxiv.org/abs/2412.16895 (accessed Jan. 25, 2025).

7 Y. Xian et al., “Exchanging-based Multimodal Fusion with Transformer,” arxiv, https://arxiv.org/pdf/2309.13340.pdf (accessed Jan. 26, 2025).

8 H. Zhang1 and T. C. Author, Towards robust multimodal sentiment analysis with incomplete data, https://arxiv.org/html/2409.20012v2 (accessed Jan. 25, 2025).

9 Arxiv, https://arxiv.org/pdf/2409.10925 (accessed Jan. 26, 2025).

10 Y. Zhou et al., “Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis,” arXiv.org, https://arxiv.org/ (accessed Jan. 25, 2025).