2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 979-8-3503-0436-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/PerComWorkshops59983.2024.10502594

# Attention-Based Fusion of Intra- and Intermodal Dynamics in Multimodal Sentiment Analysis

Ehsan Yaghoubi[1], Tuyet Kim Tran[1], Diana Borza[2], Simone Frintrop[1]
[1]Universität Hamburg
Hamburg, Germany, 22527
ehsan.yaghoubi@uni-hamburg.de
[2]Babeș Bolyai University, Cluj-Napoca, Romania

***Abstract*—Sentiment analysis, the process of predicting sentiments expressed in human communication, has evolved to multimodal sentiment analysis (MSA). Recent advances in attention-based MSA models have demonstrated the effectiveness of capturing intramodality and intermodality dynamics. However, challenges remain in achieving optimal performance and creating effective context representations across modalities. To address these challenges, we propose the Multi-Head self-attention with Context-Aware attention model, which utilizes two attention-based mechanisms to strategically capture intramodality dynamics within each modality before delving into intermodality dynamics. The experimental results on 5 datasets show the superiority of our model in comparison with the state of the arts.**

## I. Introduction

Sentiment analysis aims to extract and analyze sentiments expressed in textual chats, audio speech, visual-based interactions, or a combination of them. Traditional sentiment analysis primarily deals with text, but advancements in computational capabilities and the prevalence of audio-visual communication have led to the development of multimodal sentiment analysis (MSA). MSA enhances the understanding of human emotions and could be applicable in diverse fields such as education, customer feedback analysis, mental health monitoring, and personalized advertising.

Recent advancements in attention-based MSA models, such as the Multi-attention Recurrent Network (MARN) [17], [19], have showcased notable achievements in capturing both intramodality and intermodality dynamics. Additionally, [8] employs an LSTM-based approach to enhance context-based knowledge. Nevertheless, challenges persist in achieving optimal performance and creating effective context representations across modalities.

To address the above-mentioned weaknesses, we introduce multi-head self-attention with context-aware attention (MHCA) model, leveraging two attention-based components. The model strategically captures intramodality dynamics within each modality before delving into intermodality dynamics, ensuring a comprehensive understanding of sentiment expression in video clips. The main contribution of this work is an architecture that extracts inter-modality features before modality fusion to enhance sentiment analysis. Secondly, we conduct ablation studies to: (1) demonstrate the effectiveness of inter-modality feature extraction compared to inter-modality feature representation, and (2) highlight the impact of each modality on sentiment classification. Thirdly, we investigate the generalizability of the proposed model across 5 datasets, providing insights into its robustness. The experimental results show the competitiveness of our model in comparison with recent methods.

## II. Related work

From a high level perspective, MSA frameworks involve obtaining an accurate intra-modality repre-

sentation and finding an effective fusion method to model the inter-modality interaction [3], [22] . Modality fusion - the process of filtering and combining the features from various modalities - is at the center of MSA. [22] provides a hierarchical taxonomy of fusion methods with 8 categories.

Utterance level fusion models operate at the level of individual utterances and can be classified as early fusion, late fusion, and hybrid fusion systems. Early fusion [9], [16] combines modalities in a joint representation before classification, but such models risk over-generalization and sacrificing modality-specific nuances. Late fusion [6], [19] classifies them separately and combines results post-classification; such models are susceptible to error accumulation, increased computational complexity, and potential information redundancy. The strengths of late fusion lie in individually tuned classifiers, while early fusion saves time by training a single model. Hybrid fusion integrates both strategies using tensor fusion [17], [19]. Other methods pay more attention to the fine-grained interactions of modalities. Translation-level fusion, exemplified by MCTN [7], translates between modalities, generating an intermediate representation of common information. Word-level fusion models, like MARN [19], capture interactions over time, accommodating opposite indicators in longer sequences [15]. [13] fuses modalities at the word level, using an attention component to assess nonverbal context influence on word meaning.

Feature space manipulation fusion employs mathematical expressions, attention mechanisms, or neural networks to exploit the relationships between different modalities in the feature space [22]. In this context, [8] introduced an attention-based network that determines the importance of modalities, generating attention scores for utterances, while [5] extended this concept to Recurrent Neural Network, incorporating scaled-dot-product attention and Multi-Head Attention (MHA) [12]. [21] computes a soft-attention matrix representing interactions between modalities and uses it to weight elements in one modality. MISA [4] first projects the features into a modality-invariant and a modality-specific subspace. Finally, the projections are concatenated and a transformer-based self-attention is used for prediction. The main challenge of these works lies in providing an attention mechanism to focus on the most important part of each data modality while also understanding the relationships between different data modalities.

Despite advances, identifying effective fusion techniques and comprehending the impact of diverse modalities on predictions remains a challenge that is addressed in this work.

## III. Proposed method

Figure 1 shows the proposed Multi-head self-attention with Context-aware Attention model (MHCA). The primary objective is to predict sentiments by leveraging the unique and complementary information provided by each modality (text, audio, and video frames). This problem becomes particularly prominent when a singular modality fails to provide an accurate sentiment determination. The MHCA model effectively captures both intramodal and intermodal dynamics by employing two synergistic components: the Multi-Head Attention (MHA) module and the Context-aware Attention (CAM) module. The MHA module initially leverages self-attention mechanisms to model interactions within each modality's embeddings. The outputs of the MHA modules are then fed into Linear layers with dropout regularization to enhance the interaction within the modalities and to project the feature vectors to the same size. Then, the CAM module seamlessly integrates information across modalities, capturing the intricate intermodal connections. Finally, the outputs of three CAM modules are concatenated and passed through Linear and Average Pooling layers to predict the sentiment score values. In the remainder of this section, we detail each module of the proposed framework.

**Multi-Head Attention (MHA) Module:** The MHA module is applied to each modality separately to capture the intramodal dynamics, i.e. the important features and characteristics within a modality. It relies on self-attention (middle section of Fig.1), which allows the model to assign weights to different features

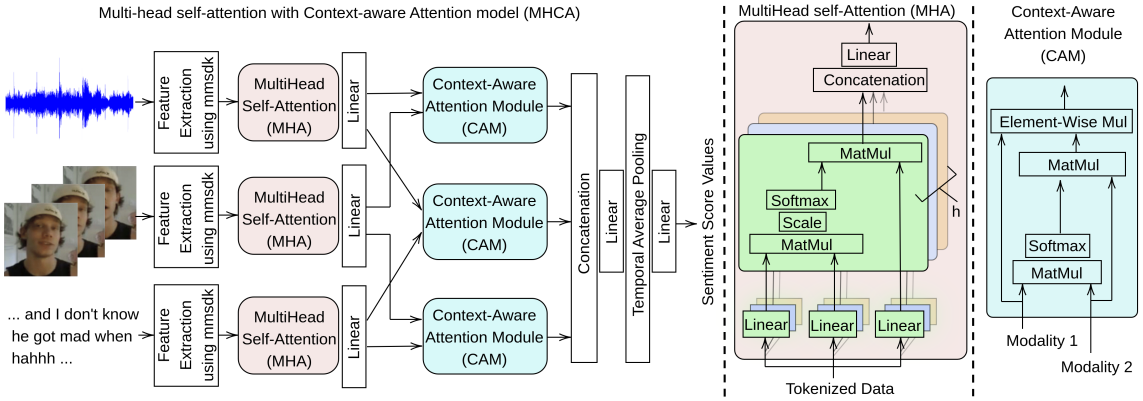


Fig. 1. Overview of the proposed model. The left side shows the overview of the model, while the middle and right sections highlight the multi-head self-attention and context-aware attention modules. *mmsdk* refers to CMU-Multimodal SDK, a toolkit for loading multimodal datasets and feature extraction. *h* stands for heads. *Mul* stands for Multiplication.

based on their contextual significance within the sequence. By using multiple self-attention heads the model has a higher capacity to capture diverse feature representations. Each head attends to specific representations within the data, allowing different important aspects within the modality to be observed. The multi-headed self-attention can be expressed as in [12]: $\mathit{MultiHead}(X) = \mathit{Concat}(\mathit{Att}(Q_1, K_1, V_1), \ldots, \mathit{Att}(Q_h, K_h, V_h))W_o$, where $\mathit{Att}(Q_i, K_i, V_i) = \zeta\left(\frac{Q_i K_i^T}{\sqrt{d}}\right)V_i$, and $Q_i = XW_{qi}$, $K_i = XW_{ki}$, and $V_i = XW_{vi}$, are respectively the query, key, and values representation of the input tokenized data $X$, and $\zeta$ is the *softmax* activation function.

**Context-aware Attention (CAM) Module:** CAM module learns the intermodality dynamics, i.e., the relationships between different modalities. The CAM module (right side of Fig. 1) uses both attention and multiplicative gating to weigh the contribution of different modalities according to their importance and context [1]: $\mathit{CAM}(M_1, M_2) = (\zeta(M_1 M_2)M_2) \odot M_1$ , where $M_1$ and $M_2$ represent the embeddings of two different modalities, $\zeta$ is the *softmax* activation function, and $\odot$ is the multiplicative gating introduced in [2].

**Modality fusion and sentiment prediction:** The outputs of the CAM modules are concatenated and fed into a Linear layer to increase interaction between modalities. Finally, temporal average pooling is used to aggregate the information across the entire sequence, and the result is passed through a Linear layer with *softmax* activation for sentiment score value prediction.

**Implementation and training:** We use the multimodal features extracted with Multimodal Software Development Kit (CMU-multimodal SDK) as inputs to our proposed architecture. the CMU-multimodal SDK is a well-known tool in the sentiment analysis domain that makes feature extraction more easily accessible and allows for fair and standardized comparisons between methods. The number of heads in the MHA module is a hyperparameter that accommodate dataset-specific features: we employed 5, 2, and 7 heads for analyzing text, audio, and visual modalities, respectively for the MMMO, Youtube, and MOUD datasets. For the MOSI dataset 5 heads were used to analyze all the modalities. Different numbers of heads are used for each dataset because the shapes of the features vary across datasets, and we should ensure that the features are divisible by the number of heads. For the final output, categorical cross-entropy guides the learning process, minimizing the discrepancy between predicted and actual sentiments. The number of neurons in the Linear layers, dropout rate, and activation function are optimized through grid search, ensuring adaptability to unique dataset characteristics [1].

[1]More details at: https://drive.google.com/file/d/1Eq6mSQDWphU3AOde_f4xfOrs7Xie49Ik/view?usp=sharing

TABLE I
PERFORMANCE COMPARISON ON THE MMMO, YOUTUBE, CMU-MOSI WITH 2 AND 7 CLASS LABELS, AND MOUD DATASETS. *VALUES ARE TAKEN FROM [11].

| Model | MMMO | | Youtube | | CMU-MOSI (2) | | CMU-MOSI (7) | | MOUD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| MV-LSTM [10]* | 72.50 | 72.30 | 45.80 | 43.30 | 73.90 | 74.00 | 33.20 | - | 57.60 | 48.20 |
| BC-LSTM [8]* | 70.00 | 70.10 | 45.00 | 45.10 | 73.90 | 73.90 | 28.70 | - | 72.90 | 72.90 |
| TFN [17]* | 72.50 | 72.60 | 45.00 | 41.00 | 74.60 | 74.50 | 28.70 | - | 63.20 | 61.70 |
| MARN [19]* | 71.30 | 70.20 | 48.30 | 44.90 | 77.10 | 77.00 | 34.70 | - | 81.10 | 81.20 |
| MFN [18]* | 73.80 | 73.10 | 51.70 | 51.60 | 77.40 | 77.30 | 34.10 | - | 81.10 | 80.40 |
| MFM [11]* | 81.30 | 79.20 | 53.30 | 52.40 | 78.10 | 78.10 | 36.20 | - | 82.10 | 81.70 |
| CIA [1] | 82.75 | 81.47 | 55.93 | **55.13** | **79.88** | 79.54 | 38.92 | - | **82.41** | **82.07** |
| R-CIA [1] | 82.50 | - | 52.54 | - | 76.68 | - | 37.46 | - | 75.47 | - |
| MHCA (ours) | **85.00** | **85.20** | **59.32** | 54.68 | 79.74 | **79.93** | **39.94** | **37.73** | 78.30 | 78.26 |

## IV. EXPERIMENTS

The evaluation of our proposed sentiment analysis models is conducted on four benchmarks: Youtube opinion [14], MMMO [19], MOUD [9], and CMU-MOSI [20] datasets. The YouTube dataset consists of 47 English-language videos containing 269 utterances about product reviews and opinions, categorized into positive, neutral, and negative sentiments. The MMMO dataset extends the Youtube dataset with 340 videos of online social reviews, labeled as positive and negative. MOUD is a collection of 79 Spanish-language videos with 389 utterances about product reviews, categorized as positive or negative. Finally, the CMU-MOSI dataset comprises 2199 opinion utterances in English sourced from YouTube, and provides both seven and two-class sentiment categories.

### A. Comparative Analysis

TableI shows the comparative results between our method and MV-LSTM [10], BC-LSTM [8], TFN [17], MARN [19], MFN [18], MFM [11], CIA [1], R-CIA [1] across four benchmarks for multimodal sentiment analysis: MMMO, Youtube, CMU-MOSI (with 2 class labels), CMU-MOSI (with 7 class labels), and MOUD. The evaluation relies on maximum accuracy and F1-score metrics. The proposed model (MHCA) outperforms all other models on three datasets, affirming its competitiveness with existing state-of-the-art models. Specifically, on the MMMO dataset, MHCA demonstrated superior accuracy, surpassing other models by margins ranging from 2.25% to 15.00%. In terms of F1-score, MHCA outperformed all models at least by 3.73%. Similarly, on the YouTube dataset, with an accuracy of 59.32%, our method surpasses the next best method CIA [1] by 3.39%. Regarding F1-score, MHCA ranked second, outperforming several models but trailing behind CIA [1] by 0.45%.

On CMU-MOSI with 2 class labels, MHCA achieved an accuracy of 79.74%, placing behind the best-performing model by 0.14%. However, in terms of F1-score, MHCA outperformed all models by 0.39%, achieving an F1-score of 79.93%. For the CMU-MOSI(7) dataset, MHCA demonstrated the highest accuracy, outperforming other models by 1.02% to 11.24%. However, F1-score comparisons were not available for other models.

On the MOUD dataset, MHCA positioned itself with accuracy of 78.30% in the middle of the compared models but showcased strong performance compared to MV-LSTM [10], BC-LSTM [8], TFN [17], and R-CIA [1]. In terms of F1-score, MHCA achieves a score of 78.26, while the best model obtains an F1-score of 82.07.

In summary, the MHCA model demonstrated exceptional performance, particularly on the MMMO, Youtube, and CMU-MOSI(7) datasets, surpassing all state-of-the-art models in the comparison. Its strong F1-score performance on the MMMO and CMU-MOSI(2) datasets further highlights its efficacy in MSA.

TABLE II
RESULT COMPARISON BETWEEN UNI-MODAL, BI-MODAL, AND TRI-MODAL EXPERIMENTS USING T (TEXT), V (VISUAL), AND A (AUDIO) MODALITIES ON THE PROPOSED MHCA MODEL.

| Model | MMMO | | Youtube | | MOSI (2) | | MOSI (7) | | MOUD | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| T | 68.75 | 70.39 | 45.18 | 52.15 | 72.59 | 73.02 | 32.80 | 25.77 | 71.70 | 72.00 |
| A | 76.25 | 77.86 | 42.37 | 0.00 | 52.92 | 59.15 | 16.47 | 0.00 | 60.37 | 74.56 |
| V | 65.00 | 72.00 | 45.76 | 0.00 | 50.87 | 55.70 | 20.85 | 0.00 | 60.37 | 75.30 |
| T+V | 71.25 | 75.33 | 50.85 | 53.41 | 77.11 | 77.23 | 36.15 | 32.83 | 73.58 | 73.70 |
| T+A | 81.25 | 81.98 | 52.54 | 51.19 | 77.11 | 77.30 | 34.55 | 34.21 | 75.47 | 75.23 |
| A+V | 70.00 | 70.00 | 49.15 | 22.38 | 58.89 | 59.29 | 23.18 | 16.04 | 61.32 | 63.72 |
| T+A+V (MHCA) | 85.00 | 85.20 | 59.32 | 54.68 | 79.74 | 79.93 | 39.94 | 37.73 | 78.30 | 78.26 |

### *B. Ablation Studies*

We conducted an ablation study to assess the effect of each module in our MHCA model. When the CAM module is removed, the model's performance experiences a notable decline. Across all datasets, the maximum accuracy values decreased by an average of 5.57%, while the mean accuracy values dropped by an average of 3.75%. This decrease indicates the role of the CAM in capturing intermodal dynamics. The most impact was observed on the MMMO dataset, where the maximum accuracy plummeted by 8.75%, reaching 76.25%. The mean accuracy also decreased by 6.50% to 72.38%, highlighting the module's pivotal role in learning intermodal dynamics. Similar trends were noted on the Youtube dataset, with an 8.47% accuracy drop to 50.85%, emphasizing the module's importance on smaller datasets.

We also examined the impact of removing the MHA module, investigating its effectiveness in capturing intramodal dynamics. When the MHA module was omitted, the model's performance was significantly compromised across all datasets. On average, the maximum accuracy values decreased by 12.25%, and the mean accuracy values dropped by an average of 7.77%.

The combined results the ablation studies highlight that excluding either the CAM or MHA module significantly hampers the model's performance. To achieve the best results, it is imperative to include both modules. This is especially crucial for achieving optimal performance on smaller datasets.

In another ablation study, we evaluated various input combinations: uni-modal (text only, audio only, video only), bi-modal (text + video, text + audio, audio + video), and tri-modal (text + audio + video) scenarios (Table II). When examining individual modalities for predicting input data sentiment, inconsistencies emerge across datasets. In certain instances, text outperforms the other two modalities, while there are situations where audio proves to be the most effective. Visual data also demonstrates superior results in some cases. Overall, text analysis tends to be more effective than audio and visual analysis, with audio and visual data showing comparable performance. In the evaluation of pairs of modalities, results surpass those of unimodal scenarios, with the combination of text consistently yielding superior results compared to the combination of audio and visual data. As anticipated, the comprehensive analysis and fusion of all three modalities lead to enhanced results compared to both bimodal and unimodal scenarios. For example, as shown in Table II, in the MMMO dataset, uni-modal analysis highlights audio's superior performance with 76.25% accuracy and 77.86% F1-score, surpassing text (68.75%, 70.39%) and video (65.00%, 72.00%). Combining text and audio in a bi-modal setup achieves the highest accuracy (81.25%, 81.98%). Tri-modal fusion excels, reaching 85.00% accuracy and 85.20% F1-score, outperforming the best bi-modal combination by 3.75% accuracy and 3.22% F1-score.

## V. CONCLUSION

In this paper, we introduced a Multi-head self-attention with Context-aware Attention (MHCA)

model for MSA and demonstrated its effectiveness in capturing both intra- and intermodality dynamics across text, audio, and video data. The multi-headed self-attention effectively captures the relationship within each modality, while the proposed CAM models the inter-modality dynamics without using any learnable parameters. The experiments on five publicly available sentiment analysis datasets demonstrated the superiority of our method in comparison with the existing methods. Further analysis of the model showed that utilizing all three modalities compared to uni-modal and bi-modal data obtains higher performance.

## ACKNOWLEDGMENT

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with Project-ID 261402652 and the National Science Foundation of China with Project-ID 62061136001 in project Crossmodal Learning, TRR-169. The work done by Diana Borza was funded by SRG-UBB 32886/21.06.2023.

## REFERENCES

[1] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya. Context-aware interactive attention for multimodal sentiment and emotion analysis. In *Proceedings of the 9th EMNLP-IJCNLP*, page 5647–5657, 2019.

[2] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, and R. Salakhutdinov. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549*, 2016.

[3] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[4] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020.

[5] T. Kim and B. Lee. Multi-attention multimodal sentiment analysis. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 436–441, 2020.

[6] Z. Pan, Z. Luo, J. Yang, and H. Li. Multi-modal attention for speech emotion recognition. In *arXiv preprint arXiv:2009.04107*, 2020.

[7] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

[8] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[9] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, 2013.

[10] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke. Extending long short-term memory for multi-view structured learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, page 338–353. Springer, 2016.

[11] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[13] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

[14] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[15] C. Xi, G. Lu, and J. Yan. Multimodal sentiment analysis based on multi-head attention mechanism. In *Proceedings of the 4th international conference on machine learning and soft computing*, pages 34–39, 2020.

[16] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021.

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[18] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[19] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] 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.

[21] S. Zhang, S. Zhang, T. Huang, and W. Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2017.

[22] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023.