Multimodal Multi-loss Fusion Network for Sentiment Analysis

Zehui Wu,[*1] **Ziwei Gong**,[* 1] **Jaywon Koo**, [1] **Julia Hirschberg** [1]
Department of Computer Science
Columbia University
{zw2804, zg2272, jk4541}@columbia.edu
julia@cs.columbia.edu

Abstract

This paper investigates the optimal selection and fusion of feature encoders across multiple modalities and combines these in one neural network to improve sentiment detection. We compare different fusion methods and examine the impact of multi-loss training within the multi-modality fusion network, identifying surprisingly important findings relating to subnet performance. We have also found that integrating context significantly enhances model performance. Our best model achieves state-of-the-art performance for three datasets (CMU-MOSI, CMU-MOSEI and CH-SIMS). These results suggest a roadmap toward an optimized feature selection and fusion approach for enhancing sentiment detection in neural networks.

1 Introduction

In recent years, the multimodal affective computing field has seen significant advances in feature extraction and multimodal fusion methodologies in recent years (Garg et al., 2022), enabling a more nuanced understanding of human emotions by effectively synthesizing audio, text, and visual signals (Sun et al., 2023a; Yu et al., 2021). This study presents a series of experiments that delve into feature selection, comparative analysis of fusion network performance, multi-loss training, and context modeling utilizing audio and text from three datasets: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018) , and CH-SIMS (Bagher Zadeh et al., 2018). Our research aims to pioneer state-of-the-art (SOTA) approaches in affective computing tasks, achieved by identifying optimal features for each modality, devising the most effective methods for their fusion, and refining training methodologies to enhance performance.

Hand-crafted feature extraction algorithms often lack flexibility and generalization across diverse tasks. To overcome these limitations, recent studies have proposed fully end-to-end models that jointly optimize feature extraction and learning processes (Dai et al., 2021; Han et al., 2021). Our work leverages feature representations derived from pre-trained models across different modalities, combining them into an end-to-end framework, which provides a comprehensive and adaptable solution for multimodal affective feature computation.

In multimodal fusion, the challenge lies in effectively fusing diverse signals, including natural language, facial gestures, and acoustic behaviors. Methods like the Tensor Fusion Network (TFN)(Zadeh et al., 2017) have been proposed to model intra-modality and inter-modality interactions. More recently, transformer encoder structures with cross-modal attention have gained popularity for integrating multimodal data (Tsai et al., 2019a; Yu et al., 2023a), with continous efforts to improve representations of multimodal information (Qian et al., 2023; Hu et al., 2022). Building from these ideas, we propose a robust fusion network structure that integrates cross-modal attention and self-attention, with additional feed-forward layers to refine the representations. Furthermore, our approach experiments with restoring original signals during the fusion process.

Our experimental analysis has identified the most efficacious features for different modalities and compared various fusion network methods for amalgamating audio and text signals. Our findings reveal that the incorporation of audio signals consistently elevates performance metrics. More notably, our transformer fusion network demonstrates a remarkable enhancement of results and and achieves state-of-the-art performances across all datasets. Additionally, our exploration into multi-loss training has yielded two significant observations: first, the utilization of distinct labels for each modality

[*] These authors contributed equally to this work.
Code is released at: https://github.com/zehuiwu/MMML

in multi-loss training markedly benefits the models' performance; second, training with multimodal features not only boosts overall model performance but also notably enhances accuracy in the single-modality subnet. Furthermore, we compared two context modeling methods and found that contextual integration significantly amplifies model performance across all metrics. These novel findings have advanced our understanding of multi-modal sentiment analysis.

2 Related Work

Existing research on multimodal affective computing often employs hand-crafted algorithms to perform initial feature representation extraction and retrieve some fixed representations for each modality (Shenoy and Sardana, 2020; Delbrouck et al., 2020). However, for these, the extracted features are static and lack the flexibility to be further fine-tuned for different target tasks; also, the manual determination of feature extraction algorithms can lead to sub-optimal performance due to constraints in generalization across diverse tasks (Dai et al., 2021). To address these issues, recent studies have proposed fully end-to-end models, effectively bridging the gap between feature extraction and learning processes (Dai et al., 2021; Wang et al., 2020). Our research also emphasizes an end-to-end structure that optimizes both phases jointly, presenting a comprehensive and adaptable solution for multimodal affective feature computation.

Lexical features, owing to pre-training on expansive corpora through Transformer-based models, often outperform other modalities. Some recent work aims to improve model performance by incorporating speech information inside the text model such as SPECTRA (Yu et al., 2023b), by pre-training a speech-text transformer model to capture the speech-text alignment effectively. A similar innovative method is the Transformer-Based Speech-Prefixed Language Model (TEASEL) (Arjmand et al., 2021), which incorporates speech as a dynamic prefix along with the textual.

Many studies have explored multimodal human language time-series data, which typically includes a mixture of natural language, facial gestures, and acoustic behaviors. However, fusing these into a unified representation presents a significant challenge due to the variable sampling rates across modalities and the difficulty in determining intra-modality dependencies. Various methods have been proposed to model the interaction across modalities (Barezi and Fung, 2019), such as the Tensor Fusion Network (Zadeh et al., 2017), which utilizes the Cartesian product of different modalities to model both intra-modality and inter-modality interactions. More recent work has shifted toward employing transformer encoder structures to integrate these signals via cross-modality attention. The MULT model (Tsai et al., 2019a) has pioneered this approach, introducing directional pairwise cross-modal attention. This method allows for interaction between multimodal sequences across distinct time steps and inherently adapts streams from one modality to another. Further research has also leveraged this concept of cross-modality attention (Goncalves and Busso, 2022; Paraskevopoulos et al., 2022), yielding valuable insights into how multimodal data can be processed more effectively. We enhance this approach by employing a self-attention encoder and a feed-forward network to further optimize the multimodal representation after one modality is projected into another using the cross-modality attention module, thus enriching our ability to process and understand multimodal data.

3 Methodology

The model for sentiment detection in our study involves two primary components: the feature network and the fusion network. Each of these has its own unique mechanisms and contributes towards the overall functioning of our proposed Multi-Modality Multi-Loss Fusion Network (MMML) as illustrated in Figure 1. Additionally, We adopted multi-loss training, experimented with restoring original signals, and explored context modeling.

3.1 Feature Network

The Feature Network employs two different pre-trained models for text and audio processing. The text subnet leverages RoBERTa (Liu et al., 2019), chosen for its significantly superior performance on various downstream tasks. The audio subnet employs different models for different languages: HuBERT (Hsu et al., 2021) for Mandarin and Data2Vec (Baevski et al., 2022) for English. This ensures the optimized extraction of features from the given modalities, setting a solid foundation for the subsequent fusion process.

The details and results of the feature network selection process are in Appendix A.5, and the

Figure 1: Our Model Structure

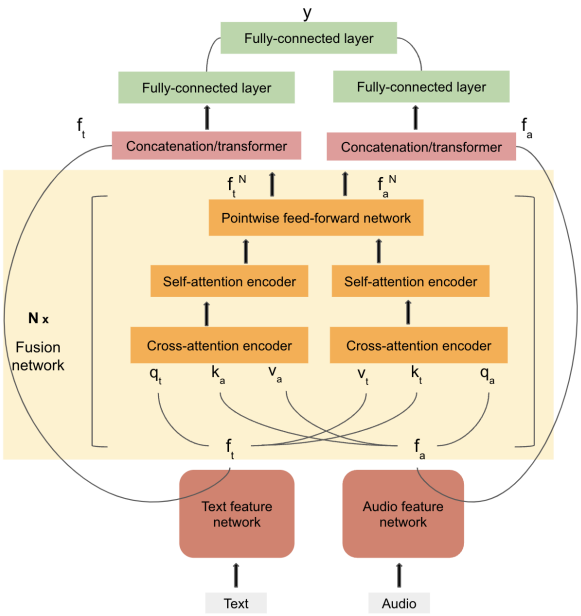

Figure 2: Model Variations

experiments that lead to the exclusion of the vision modality are in Appendix A.3 and A.4.

3.2 Fusion Network

The Fusion Network is the heart of the MMML, where the information from multiple modalities is combined. This network is divided into three smaller components as shown in the yellow portion of Figure 1.

First, there is a Cross-Attention Encoder, which adopts a mechanism similar to the self-attention encoder but which employs a query from one modality and uses keys and values generated from another modality. This cross-modal interaction aims to capture the inter-dependencies between different modalities, contributing to a more holistic understanding of the data. This encoder is defined as:

$$\text{Attention}(Q_{m1}, K_{m2}, V_{m2}) = \text{softmax}\left(\frac{Q_{m1}K_{m2}^T}{\sqrt{d_k}}\right)V_{m2}$$

where:

$$Q_{m1} = W_q \cdot f_{m1}$$
$$K_{m2} = W_k \cdot f_{m2}$$
$$V_{m2} = W_v \cdot f_{m2}$$

We denote queries as Q_{m1} (from modality 1) and the keys and values as K_{m2} and V_{m2} (from modality 2). f_{m1} is the feature from modality 1 and f_{m2} is the feature from modality 2. In Figure 1, the text feature f_t is utilized to build the query q_t for text and is also used to construct the key k_a and value v_a for audio. The cross-attention encoder essentially projects the hidden states from one modality into the space of another modality.

Also, our proposed network includes additional *Self-Attention Encoders*. The self-attention mechanism was originally designed to find the correlation within a single modality, thereby capturing the intra-modal dynamics of the data. In our model, the self-attention module serves to model the connections across time steps of the new feature representation after passing through the cross-modality encoder.

Finally, our model includes a *Pointwise Feed-Forward Network* which applies fully connected feed-forward networks and ReLU activation functions to each individual position, further refining the encoded feature representations. Through combining these methodologies, we aim to optimize the multi-modal feature extraction and fusion process, enhancing the MMML's performance in sentiment detection tasks.

3.3 Multi-Loss Training

In order to leverage *multi-loss training*, we modified the architecture of our fusion network to incorporate an additional fully-connected layer at the termination of each feature network, as illustrated in the green portion of Figure 1. This modification enables two additional outputs from individual modalities, in addition to the combined feature

output. This design facilitates the application of three distinct loss functions during training, each corresponding to one of the outputs. Despite the classification being conducted using only the output from the fusion network, the final loss is the summation of losses from the subnets and the loss from the fusion network:

$$Loss = \sum_{m \in \{a,t,f\}} \alpha_m * loss\_fn(y_m, target_m)$$

The source of targets from different modalities is explained in section 4.5 and we witness no significant boost in performance by adjusting α to be other than 1.

The rationale behind implementing additional losses for each individual modality is to bolster the respective feature networks' comprehension and processing of their respective signals. Given that each feature network perceives and handles signals distinctively, akin to how humans discern emotions through different sensory signals, the multi-task loss serves a dual purpose: first, it encourages each feature network to refine its method of processing its specific modality, akin to honing the "sense" associated with that modality; second, it trains the fusion network to effectively combine the distinct signals relayed by the feature networks, as guided by the loss from the combined modality.

Through this multi-loss training approach, we create a model that efficiently mirrors human-like multi-modal emotion perception, each modality working independently and collaboratively to understand the comprehensive emotional context.

3.4 Original Signal Restoration

In our investigation of the fusion network, we developed two variants designed to mitigate potential loss of original signals during the cross-modal projection process, as shown in Figure 2. Because the cross-attention mechanism projects one modality into another, some original signals might be obscured or lost. Therefore, these variations aim to combine the original signal with the projected signal, thereby enhancing the ability of the network to learn from both signals simultaneously.

The first is *Concatenation Variation*, which concatenates the original feature with the fused feature. The second is the *Transformer Variation*. This variation merges the original hidden states and the fused hidden states along the feature dimension and uses transformer encoders to further process these combined hidden states.

This fusion of original and projected information within each modality aims to maintain the integrity of the original signals, while also integrating the enriched cross-modal information. The combined features then go through a linear layer and are subsequently concatenated with features from other modalities.

3.5 Context Modeling

We explored the integration of contextual data (previous utterances) into existing model frameworks, specifically contrasting two distinct methodologies for context integration. These methodologies were: (i) the **concatenation** of context (previous utterances) and the current utterance as a singular input stream to the model, and (ii) **independent processing** of context (previous utterances) and current utterance, followed by a subsequent fusion of their respective representational outputs. The first method treats the concatenated input as a single utterance and gives one representation, while the second processes the context and current input separately and gives one representation for each.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We use three primary datasets, each characterized by its unique properties and content, to test the performance of the Multi-Modality Multi-Loss Fusion Network on sentiment detection.

The *CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI)* (Zadeh et al., 2016): This dataset, developed in English, includes audio, text, and video modalities compiled from 2199 annotated video segments collected from YouTube monologue movie reviews. It offers a focused approach to studying sentiment detection within the context of film critiques.

The *CMU-Multimodal Sentiment Analysis (CMU-MOSEI)* (Bagher Zadeh et al., 2018) is an extension of CMU-MOSI, incorporating the same modalities of audio, text, and video from YouTube videos, but it has a broader scope, covering a wider range of topics, and is more substantial in size, with 23,453 annotated video segments.

The *Chinese Multimodal Sentiment Analysis Dataset (CH-SIMS)* (Yu et al., 2020) includes the same modalities in Mandarin: audio, text, and video, collected from 2281 annotated video segments. It includes data from TV shows and movies,

Model	CMU-MOSI							CMU-MOSEI						
	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr
LMF	-	-	82.5	82.4	33.20	0.917	0.695	-	-	82.0	82.1	48.00	0.623	0.700
TFN	-	-	80.8	80.7	34.90	0.901	0.698	-	-	82.5	82.1	50.20	0.593	0.677
MFM	-	-	81.7	81.6	35.40	0.877	0.706	-	-	84.4	84.3	51.30	0.568	0.703
MTAG	-	-	82.3	82.1	38.90	0.866	0.722	-	-	-	-	-	-	-
SPC	-	-	82.8	82.9	-	-	-	-	-	82.6	82.8	-	-	-
ICCN	-	-	83.0	83.0	39.00	0.862	0.714	-	-	84.2	84.2	51.60	0.565	0.704
MulT	81.50	80.60	84.10	83.90	-	0.861	0.711	-	-	82.5	82.3	-	0.580	0.713
MISA	80.79	80.77	82.10	82.03	-	0.804	0.764	82.59	82.67	84.23	83.97	-	0.568	0.717
COGMEN	-	-		84.34	43.90	-	-	-	-	-	-	-	-	-
Self-MM	84.00	84.42	85.98	85.95	-	0.713	0.798	82.81	82.53	85.17	85.30	-	0.530	0.765
MAGBERT	84.20	84.10	86.10	86.00	-	0.712	0.796	84.70	84.50	-	-	-	-	-
MIMM	84.14	84.00	86.06	85.98	46.65	0.700	0.800	82.24	82.66 4	85.97	85.94	54.24	0.526	0.772
TEASEL	84.79	84.72	87.5	85	47.52	64.4	83.6	-	-	-	-	-	-	-
SPECTRA	-	-	87.5	-	-	-	-	-	-	87.34	-	-	-	-
UniMSE	85.85	85.83	86.9	86.42	48.68	69.1	80.9	85.86	85.79	87.5	87.46	54.39	52.3	77.3
MMML	85.91	85.85	88.16	88.15	48.25	64.29	83.8	86.32	86.23	86.73	86.49	54.95	51.74	79.08
+ context	**87.51**	**87.45**	**89.69**	**89.67**	**50.34**	**58.31**	**86.93**	**87.24**	**87.18**	**88.02**	**88.15**	**55.74**	**49.22**	**81.37**

(a) CMU-MOSI and CMU-MOSEI

	ACC_2	ACC_3	ACC_5	F1	MAE	Corr
EMT	80.1	67.4	43.5	80.1	39.6	62.3
MMML(ours)	**82.93**	**69.37**	**49.38**	**82.9**	**33.2**	**73.26**

(b) CH-SIMS

Table 1: **Comparison with SOTA**: Achieved best performance on three datasets. All experimental results presented are averages derived from three separate runs. The performances of baselines are shared by their authors.

making it culturally distinct and diverse, and provides multiple labels for the same utterance based on different modalities, which adds an extra layer of complexity and richness to the data.

These datasets provide a broad and multicultural perspective on sentiment detection, allowing for a thorough evaluation and comparative analysis of the MMML's performance across diverse data landscapes.

4.1.2 Baseline Models

In our comprehensive evaluation of the MMML model, we conducted a detailed comparison with a wide array of baseline models in multimodal sentiment analysis. This comparison included several categories of models, each representing a unique approach to multimodal learning.

The first category consisted of early multimodal fusion methods, including Tensor Fusion Network (TFN) (Zadeh et al., 2017), Low-rank Multimodal Fusion (LMF) (Liu et al., 2018), and Multimodal Factorization Model (MFM) (Tsai et al., 2019c). These models are foundational in early multimodal fusion approaches to multimodal analysis.

The second category focused on methods that enhance multimodal integration through more modern modality interaction modeling methods. This includes the multimodal Transformer (MulT) by Tsai et al. (2019b), Interaction Canonical Correlation Network (ICCN) by Sun et al. (2019), Sparse Phased Transformer (SPC) by Chen et al. (2021), and Modal-Temporal Attention Graph (MTAG) by Yang et al. (2021). These models represent a significant advancement in handling complex multimodal interactions.

The third category encompasses models that prioritize modality consistency and difference. This includes MISA (Hazarika et al., 2020), which manages modal representation space, Self-MM (Yu et al., 2021) which leverages multi-task learning from unimodal representations, MAG-BERT (Rahman et al., 2020) with its innovative fusion gate, and MMIM (Han et al., 2021) which focuses on maximizing mutual information hierarchically. We also evaluated models focusing on self-supervised learning Transformers for combined modalities, such as TEASEL (Arjmand et al., 2021), and those exploring speech-text alignment like SPECTRA (Yu et al., 2023b).

Further broadening our comparison, we included models like COGMEN (Joshi et al., 2022), which consider the multimodal conversational context. A noteworthy inclusion was UniMSE (Hu et al., 2022), a model proposing a knowledge-sharing framework that unifies multimodal sentiment analysis and emotion recognition in conversation tasks, currently regarded as the state-of-the-art (SOTA) method.

In comparing the feature extractor between ours

and the baselines, we find our approach to be on par with the baselines in terms of selecting feature extractors. Like ours, the majority of the baseline models employ advanced pre-trained models for feature extraction. For example, TEASEL adds speech tokens on top of a pre-trained RoBerta during training, which also uses advanced pre-trained models as the feature extractor, which is comparable to ours.

Our methods further explore the effectiveness of cross-modality attention, aiming to further optimize the multimodal representation during the projection of one modality into another, along with enhancing to select optimal training methods including multi-loss training and context-modeling, thus enriching our ability to process and understand multimodal data.

4.1.3 Metrics

Our MMML model was evaluated using metrics consistent with existing research against existing benchmarks, which enables comprehensive evaluation of our model's performance across diverse sentiment analysis dimensions (detailed descriptions in Appendix A.6). Additional sets of ablation experiments on different components of the model were conducted for analysis, to interpret and explain the model performance.

4.2 Results

Our overall results are shown in Table 1. When compared with contemporary state-of-the-art models, our method emerges as a robust performer, offering superior outcomes for both CMU-MOSI and CMU-MOSEI. Among recent models, UniMSE (Hu et al., 2022) has delivered the best results on the English datasets. Nonetheless, our MMML model surpasses UniMSE in most of the evaluation metrics, reinforcing the effectiveness of our approach. Adding context further elevates performance substantially for all metrics.

For CH-SIMS, one state-of-the-art model is the Efficient Multimodal Transformer (EMT) (Sun et al., 2023b), which has demonstrated a high degree of performance over existing methods. Our MMML model also significantly outperforms EMT across all metrics, further underscoring the potential of our fusion network.

We also note that our model uses only audio and text signals, while these other models take advantage of all three signals. This further proves the effectiveness of our model. These results not only validate our multimodal fusion network but also affirm the robustness of our chosen methodology for sentiment detection tasks. Our impressive performance on all three datasets, CMU-MOSI, CMU-MOSEI, and CH-SIMS, verifies the versatility and adaptability of our MMML model, emphasizing its value in advancing the field of sentiment detection.

Feature name	$\mathbf{ACC_2}$
openSMILE	0.6696
Mel Spectrogram	0.6805
Fine-tuned HuBert(CH)	**0.7465**

(a) CH-SIMS

Feature name	$\mathbf{ACC_2}$
openSMILE	0.4606
Mel Spectrogram	0.4519
Fine-tuned Data2vec(EN)	**0.7099**

(b) CMU-MOSI

Table 2: **Audio Feature Selection Results**: Fine-tuning a pre-trained audio model works significantly better than using other audio features.

4.3 Audio Feature Comparison

To incorporate the best speech information into our model, the initial stage of our experimentation process involved comparing the performance on audio features for sentiment analysis from two datasets, CMU-MOSI (English) and CH-SIMS (Mandarin), using openSMILE and Mel spectrograms, each with customized parameters for optimal feature extraction to compare with features from a pre-trained audio model (details in Appendix A.2).

Upon evaluation of the different audio feature extraction methods shown in Table 2, we found that use of a pre-trained model for raw audio yielded higher accuracy rates: accuracy rates of approximately 71% and 75% were achieved for CMU-MOSI and CH-SIMS, respectively. This outperformed the other two hand-crafted features (openSMILE and Mel spectrograms) by a significant margin. Interestingly, openSMILE and Mel spectrograms displayed comparable performance on CH-SIMS. However, their performance on CMU-MOSI was notably subpar. We hypothesize that CH-SIMS, comprising audio from TV shows and movies, presents a more straightforward task for audio sentiment detection.

This analysis highlights the effectiveness of using pre-trained models for raw audio in achieving superior sentiment classification accuracy. It also underscores the need to consider the characteris-

tics and source of audio data in applying different feature extraction techniques.

4.4 Fusion Network Ablation Experiment

In our ablation study focusing on the fusion network with the CMU-MOSI dataset, we identified crucial components that significantly contribute to the network's performance. Specifically, the self-attention layers and the three fully connected layers following the cross-attention layers proved to be vital, as demonstrated in Table 8 of Appendix A.7.

Upon removing the self-attention layers, we observed a noticeable decline in model performance across all metrics. This decline became more pronounced when both the self-attention and fully connected layers were eliminated. These findings underscore the importance of these components in enhancing and refining modality representations within the fusion network, highlighting their integral role in our model's architecture.

4.5 Comparison of Simple Concatenation and Fusion Network

To demonstrate the superiority of our proposed fusion network, we compared it to concatenation. Upon analysis of our results (Table 3), we observed that the introduction of the transformer fusion network yielded improvements in performance in most metrics for CMU-MOSEI and CH-SIMS, and for half of the metrics for CMU-MOSI. These results underscore the effectiveness of our transformer fusion network in enhancing cross-modality modeling and suggest its potential as a powerful tool for multi-modal sentiment detection.

Beyond these observations, it is important to highlight that both methods which combined audio and text signals outperformed methods utilizing only text signals in all metrics across the three datasets. A noteworthy increase in performance was recorded on the CH-SIMS dataset upon the addition of audio signals, while the two English datasets, CMU-MOSI and CMU-MOSEI, exhibited smaller improvements. The substantial improvement observed in CH-SIMS can be attributed to two factors. First, CH-SIMS assigns unique labels to audio and text, thereby facilitating the network's ability to learn distinct signals from each modality. Second, the source for CH-SIMS is TV show and movie videos, which typically display easily-interpretable sentiments. This characteristic probably contributes to the effectiveness of combining audio and text signals for sentiment detection.

4.6 Multi-loss Training Experiments

To investigate the effectiveness of multi-loss training, we performed comparative experiments on two different datasets: CMU-MOSEI and CH-SIMS. CMU-MOSEI provides a single target for each utterance, whereas CH-SIMS offers different labels for each modality in addition to the combined modalities. We easily adapted multi-loss training to CH-SIMS, given its distinct labels for each modality. For CMU-MOSEI, we duplicated the single target across different losses to enable multi-task training.

The results, as shown in Table 4, were striking: while multi-loss and single-loss training performed similarly on CMU-MOSEI, multi-loss training significantly boosted performance on CH-SIMS. This underscores the value of unique labels for each modality when employing multi-loss training. The improved performance on CH-SIMS can be attributed to the distinct nature of the signals processed by each feature network. Since audio includes acoustic signals that are not present in the text, it is common for them to produce different sentiments. Having distinct labels assists each network in learning better how to process its unique signal.

Surprisingly, as shown in Table 5, the multi-loss training also contributed to an enhanced performance of the text subnet when compared to training with only the text. The additional audio signal appears to support the performance improvement. This suggests that even when the goal is to use only the text input for inference, multi-loss training can be beneficial. The text subnet, after training with the multi-modal model, can be extracted and used independently, offering superior performance compared to being trained alone.

Interestingly, this improvement was not observed in the audio subnet, potentially due to the stronger signal from the text subnet (reflected by a 10% higher accuracy when trained alone) which made it easier to train, and thus the network might have focused on reducing its loss.

In summary, the benefits of multi-loss training are threefold. First, it substantially boosts the performance of the entire network when distinct labels for different modalities are available. Second, loss from other modalities enhances the performance of the text subnet, indicating that we can utilize other modalities in training even when the text subnet is the only required component for inference.

Third, it is capable of handling missing modalities, enabling outputs when only text or audio inputs are available. These findings shed light on the potential of multi-loss training in the context of multi-modality fusion networks, opening avenues for further research and optimization.

4.7 Results for Original Signal Restoration

To understand the effect of restoring original signals, we conducted a comparative analysis of proposed fusion network variations, which reveals a relatively consistent performance across all variations. As shown by the results presented in Table 6 in Appendix A.7, the three methods demonstrate similar performance across all metrics for both CMU-MOSEI and CH-SIMS. Surprisingly, re-incorporating the original signal into the fused signal did not lead to any significant improvement in performance. In essence, while similar performance across different fusion network variations was unanticipated, it paves the way for a deeper understanding of the interactions within the fusion network and the role of original signals in such approaches.

4.8 Context Modeling Experiments

In our investigation, we sought to explore the integration of contextual information utilizing the CMU-MOSI dataset, characterized by many sequential utterances in dialogues, in contrast to the CH-SIMS dataset.

This exploration was initially conducted utilizing solely textual data streams to compare the two methods ((i) concatenation and (ii) independent processing), with different contextual window lengths (i.e., the number of preceding utterances considered as contextual input). As shown in Table 7 of Appendix A.7, our empirical findings indicated a superior capacity for context window management in the (ii) methodology that processes context and current utterance separately. Notably, while the optimal performance of the (i) concatenation approach was observed at a window of 1, the (ii) independent processing method exhibited performance enhancements up to a context window of 2. In addition, the (ii) independent processing method demonstrated superior performance across all evaluated metrics.

We next extended our research to the incorporation of audio signals, examining the effect of various permutations of context window lengths of both text and audio inputs. Optimal results were achieved with a textual context window of 2 and an auditory context window of 1. As shown in Table 1, this contextual model markedly outperformed our best non-contextual model across all evaluative metrics, identifying a significant advancement in the model's performance capabilities.

5 Conclusion

In conclusion, this study has provided novel, important findings for multi-modal sentiment analysis that should benefit future researchers in the designing of sentiment analysis and other models. First, the use of pre-trained models for raw audio yielded superior results, highlighting their effectiveness in feature extraction. Second, combining audio and text signals consistently outperformed using text signals alone, with the transformer fusion network showing promise in enhancing cross-modality modeling. Third, multi-loss training proved beneficial for performance, particularly with unique labels for each modality. Fourth, context information boosts model performances significantly. Last, achieving state-of-the-art results on three sentiment detection datasets underscores the effectiveness of our approach.

Moreover, in analysis of ablation studies, we show model performance can be improved through a method that reflects a similar pattern with human understanding of sentiment analysis through the multi-loss training. Moreover, multimodal features improve both the overall model performance in multimodality and in single-modality input settings. We provide more effective ways to handle missing modalities and utilized individual modality representation, by allowing the model to train on multimodal features and boost performance on single modality input using multi-loss training. Still, the performance of fusion network variations did remain consistent, prompting further investigation.

Acknowledgements

This research is supported in part by the Defense Advanced Research Projects Agency (DARPA), via the CCU Program contract HR001122C0034. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

	$\mathbf{Has0\_ACC_2}$	Has0_F1	$\mathbf{Non0\_ACC_2}$	Non0_F1	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	MAE	Corr
text-only	84.89	84.86	87.04	87.07	54.81	47.32	66.62	83.27
concatenation	85.77	85.74	87.6	87.62	**56.51**	**48.79**	**64.27**	**84.06**
+ fusion network	**85.91**	**85.85**	**88.16**	**88.15**	56.08	48.25	64.29	83.8

(a) CMU-MOSI

	$\mathbf{Has0\_ACC_2}$	Has0_F1	$\mathbf{Non0\_ACC_2}$	Non0_F1	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	MAE	Corr
text-only	84.81	84.95	86.34	86.19	54.99	52.7	53.31	78.6
concatenation	84.77	84.9	**86.82**	**86.65**	55.99	53.94	51.63	**79.81**
+ fusion network	**86.32**	**86.23**	86.73	86.49	**57.32**	**54.95**	**51.54**	79.08

(b) CMU-MOSEI

	$\mathbf{ACC_2}$	$\mathbf{ACC_3}$	$\mathbf{ACC_5}$	F1	MAE	Corr
text-only	79.21	65.06	42.02	79.14	42.65	59.4
concatenation	81.91	**70.68**	47.12	82.1	34.96	72.37
+ fusion network	**82.93**	69.37	**49.38**	**82.9**	**33.2**	**73.26**

(c) CH-SIMS

Table 3: **Concatenation vs. Transformer Fusion**: Integration of audio signals enhances performance across almost all metrics, with more pronounced impact on CH-SIMS. Implementing the Fusion Network augments performance slightly in most metrics. All experimental results presented are averages derived from three separate runs.

	$\mathbf{Has0\_ACC_2}$	Has0_F1	$\mathbf{Non0\_ACC_2}$	Non0_F1	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	MAE	Corr
single-loss	**85.22**	**85.39**	**87.02**	**86.91**	55.95	53.85	51.96	79.68
multi-loss	84.77	84.9	86.82	86.65	**55.99**	**53.94**	**51.63**	**79.81**

(a) CMU-MOSEI

	$\mathbf{ACC_2}$	$\mathbf{ACC_3}$	$\mathbf{ACC_5}$	F1	MAE	Corr
Single-loss	78.34	67.18	46.83	78.59	39.09	62.69
multi-loss	**81.91**	**70.68**	**47.12**	**82.1**	**34.96**	**72.37**

(b) CH-SIMS

Table 4: **Single-Loss Training vs. Multi-Loss Training**: While multi-loss training does not yield performance improvement when identical labels are used for different losses, as in the case of CMU-MOSEI, it does contribute significantly to performance enhancement when unique labels are assigned to each modality, as observed with CH-SIMS.

	$\mathbf{Has0\_ACC_2}$	Has0_F1	$\mathbf{Non0\_ACC_2}$	Non0_F1	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	MAE	Corr
text-loss only	**84.81**	**84.95**	86.34	86.19	54.99	52.97	53.31	78.6
multi-loss	84.36	84.62	**86.85**	**86.76**	**56.06**	**53.61**	**52.35**	**79.49**

(a) CMU-MOSEI

	$\mathbf{ACC_2}$	$\mathbf{ACC_3}$	$\mathbf{ACC_5}$	F1	MAE	Corr
text-loss only	79.21	65.06	42.02	79.14	42.65	59.4
multi-loss	**83.15**	**72.14**	**48.21**	**83.74**	**28.58**	**78.72**

(b) CH-SIMS

Table 5: **Impact of Multi-Loss on Text Subnet**: Utilizing audio-related losses can enhance performance of the text subnet, even when identical labels are employed, as is the case with CMU-MOSEI. Remarkably, using specific labels for different modalities results in a substantial performance boost in the text subnet, as evidenced by the results from CH-SIMS.

6 limitations

One limitation of this paper is that the proposed model is studied on two language, English and Mandarin. Another is that the coverage of domains is limited to the design of the datasets we choose to use, which is from YouTube Videos and TV shows. Hence, it is likely that a portion of the data is acted rather than naturally occurring in real life, and acted emotions may be expressed differently than naturally occurring emotions. Such bias in the dataset might lead to learning similar bias in model features and cause errors in recognition if applied to real life situations, which could be different in characteristics and distribution than YouTube videos and TV shows.

Another limitation is that among the 3 public dataset we used, which are all collected from YouTube and TV shows, not all have detailed descriptions about anonymization of the persons appeared in the dataset. However, we did not modify the dataset, since the datasets are widely used and we would like to create coherent results, comparable with previous work.

As for potential risk of misuse, since the paper is focused on more fundamental aspects of the research, it is possible that the model might not perform well if deployed in other scenarios without additional fine-tuning and training, because the model is trained on public datasets collected from TV shows and YouTube. Misuse of directly deploying the model in real-life applications might create risks that the prediction might not always be accurate.

Finally, our model does not yet incorporate vision features. In developing the model, we initially experimented with incorporating vision features including openFace features and embeddings from a finetuned VGGFace2 model during the early stages. However, our findings indicated that these features did not significantly enhance performance and required a substantial increase in computational resources — without a commensurate improvement in results. Importantly, our current model achieves performance metrics that surpass state-of-the-art (SOTA) models which do incorporate vision features. This accomplishment underscores the effectiveness of our approach relying on audio and text. Nonetheless, integrating vision features remains a potential avenue for future development. The exploration of vision capabilities is a promising direction for enhancing our model's performance, particularly in areas where visual context can provide additional insights.

References

Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. Teasel: A transformer-based speech-prefixed language model.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Elham J. Barezi and Pascale Fung. 2019. Modality-based factorization for multimodal fusion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 260–269, Florence, Italy. Association for Computational Linguistics.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?

Haoxing Chen, Huaxiong Li, Yaohui Li, and Chunlin Chen. 2021. Sparse spatial transformers for few-shot learning. *CoRR*, abs/2109.12932.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.

Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. 2022. Multimodality for NLP-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847, Marseille, France. European Language Resources Association.

Lucas Goncalves and Carlos Busso. 2022. Robust audiovisual emotion recognition: Aligning modalities,

capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. *arXiv preprint arXiv:2005.03545*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Georgios Paraskevopoulos, Efthymios Georgiou, and Alexandros Potamianos. 2022. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12966–12978, Toronto, Canada. Association for Computational Linguistics.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Association for Computational Linguistics.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023a. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 658–670, Toronto, Canada. Association for Computational Linguistics.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023b. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–17.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *CoRR*, abs/1911.05544.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019c. Learning factorized multimodal representations. In *ICLR*.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, WWW '20, page 2514–2520, New York, NY, USA. Association for Computing Machinery.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021, Online. Association for Computational Linguistics.

Jianfei Yu, Kai Chen, and Rui Xia. 2023a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1966–1978.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023b. Speech-text dialog pre-training for spoken dialog understanding with explicit cross-modal alignment.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Yuan Ziqi, and Wu Jiele. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos.

A Appendix

A.1 Training Details

The training process employed a learning rate of 1e-5, batch size of 16, and the AdamW optimizer. L2 loss was used to optimize the model during the training process. The validation set loss and accuracy were monitored to ensure that the model was not overfitting to the training data. An early stopping mechanism with patience of 8 epochs was employed to ensure the generalizability of the model. The entire procedure was conducted on a single RTX 4090 GPU. For the audio pre-trained model, the Convolutional Neural Network (CNN) portion used for feature extraction was frozen. The impact of different learning rates for various parts of the network was explored, but no significant differences were observed. Moreover, we found that using 5 layers of fusion network achieves the best results. All the results presented in the tables are averaged over three independent runs.

A.2 Audio Feature Extraction and Modeling Details

For openSMILE, we manipulated the frame size and step, setting them to 0.06 seconds and 0.02 seconds respectively. For Mel spectrograms, the number of Mel filterbanks was set to 128, while the window size and step were adjusted to 0.06 seconds and 0.02 seconds respectively. These configurations were chosen to enhance the precision of audio feature extraction without sacrificing computational efficiency.

Following the feature extraction phase, these features were used to construct models with varying architectures: Transformer models, incorporating between 2 and 4 encoder layers complemented with positional encoding, were employed to process the openSMILE features. A feed-forward layer was subsequently added to process feature embeddings in the CLS token of the final transformer encoder. For processing Mel spectrogram features, we leveraged convolutional neural network (CNN) models, including a custom 8-layer CNN model and modified versions of ResNet-18 and ResNet-32. The choice of these CNN architectures was driven by their known effectiveness in handling image-like data structures such as spectrograms.

A.3 Vision Features Experiments

For facial extraction, we initially employed MTCNN, but switched to the OpenCV DNN model due to its superior performance under dim lighting conditions. We extracted faces at a rate of 5 frames per second.

Our experiments involved two models: 1) a CNN-transformer model, combining a CNN pre-trained on the VGG-face2 dataset for spatial relationships with a transformer for temporal relationships; 2) TimesFormer (Bertasius et al., 2021), a transformer-based model pre-trained on video data. Both models achieved about 70% accuracy on the CH-SIMS dataset (0.7045 for CNN-transformer and 0.7294 for TimesFormer). However, their performance on CMU-MOSI was significantly lower (40-50% accuracy), leading us to conclude that the facial expressions in CMU-MOSI lack the distinct emotional information necessary for effective sentiment analysis.

A.4 Modality Selection

Our model's foundation is a text-based Large Language Model (LLM), while the other two modalities serve as auxiliary signals. Standalone, the text LLM, exhibits over 84 percent accuracy in English datasets and nearly 80 percent accuracy in the Mandarin dataset, outperforming other modalities in all evaluated metrics. We explored combinations of text with audio or video. The text-audio combination yielded a significant performance boost (see Table 3), whereas the text-video combination was less consistently beneficial. Specifically, video signals provide slight performance improvement in the CH-SIMS dataset, but not in the CMU-MOSI dataset. Our analysis revealed that CMU-MOSI videos contain more restrained and ambiguous facial expressions, making them less useful for sentiment detection compared to the emotionally richer expressions in CH-SIMS, sourced from TV shows and movies.

Given these findings, we chose not to include vision features in our final model. This decision was driven by two factors: 1) the inconsistent performance benefits, dependent on the dataset; 2) the substantial computational resources required for processing visual data (details explained in Appendix A.3). This latter aspect not only increases computational overhead but also introduces delays in real-time inference, without a corresponding improvement in results. We recognize that our approach lacks scalability in integrating multiple modalities due to the necessity of large pre-trained models for each to obtain the best performance.

Importantly, our model still surpasses state-of-the-art models that include vision features in performance metrics, emphasizing the efficacy of our text and audio-based approach. Nevertheless, the potential for integrating vision features remains a promising area for future enhancements. The utility of visual context, especially in scenarios where it provides critical insights, warrants further exploration.

A.5 Feature Network Selection

In our sentiment analysis framework, we conducted extensive experiments with various text and audio models. For text processing, we evaluated MacBert, Bert, and RoBerta across both Mandarin and English datasets. RoBerta consistently outperformed the others in both languages, demonstrating its superior efficacy in understanding and processing textual data.

For the audio component in the English dataset, we tested three Automatic Speech Recognition (ASR) models: wave2vec, HuBert, and Data2Vec. Among these, Data2Vec emerged as the most effective, providing the best performance in terms of accuracy and reliability. As there was no fine-tuned version of Data2Vec for Mandarin that is publicly available, we compared the performance of wave2vec and HuBert, ultimately selecting HuBert for its superior performance with Mandarin. An interesting observation from our experiments was the positive correlation between the performance of ASR models in sentiment detection and their word error rate (WER) in public benchmarks.

A.6 Metrics

Our MMML model was evaluated using metrics consistent with existing research.

For CMU-MOSI and CMU-MOSEI, we used:

- $\mathbf{Has0\_ACC_2}$, $\mathbf{Has0\_F1}$, including zero sentiment scores as positive;
- $\mathbf{Non0\_ACC_2}$, $\mathbf{Non0\_F1}$, ignoring zero sentiment scores;
- $\mathbf{ACC_5}$, $\mathbf{ACC_7}$, representing 5-class and 7-class accuracies respectively;
- $\mathbf{MAE}$, Mean Absolute Error;
- $\mathbf{Corr}$, assessing the correlation between predicted and actual scores.

For CH-SIMS, we utilized:

- $\mathbf{ACC_2}$, $\mathbf{ACC_3}$, $\mathbf{ACC_5}$, representing 2-class, 3-class, and 5-class accuracies respectively;
- $\mathbf{F1}$, balancing precision and recall;
- $\mathbf{MAE}$, mean absolute error;
- $\mathbf{Corr}$, assessing correlation between predicted and actual scores.

These metrics enable comprehensive evaluation of our model's performance across diverse sentiment analysis dimensions.

A.7 Additional tables

	$\mathbf{Has0\_ACC_2}$	$\mathbf{Has0\_F1}$	$\mathbf{Non0\_ACC_2}$	$\mathbf{Non0\_F1}$	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	$\mathbf{MAE}$	$\mathbf{Corr}$
Fused Features Only	**86.32**	**86.23**	86.73	86.49	**57.32**	54.95	**51.54**	79.08
Concatenation	84.96	85.09	**86.78**	**86.61**	56.86	**57.78**	51.88	**79.09**
Transformer	86.11	86.08	86.7	86.46	57.01	54.31	51.97	78.96

(a) CMU-MOSEI

	$\mathbf{ACC_2}$	$\mathbf{ACC_3}$	$\mathbf{ACC_5}$	$\mathbf{F1}$	$\mathbf{MAE}$	$\mathbf{Corr}$
Fused Features Only	**82.93**	69.37	49.38	**82.9**	33.2	**73.26**
Concatenation	82.42	69.44	49.82	82.38	33.6	72.87
Transformer	82.42	**69.95**	**49.89**	82.52	**33.12**	72.61

(b) CH-SIMS

Table 6: **Comparative Performance of Model Variations**: The *Fused Features Only* model employs only the features following the fusion network, while the *Concatenation* model merges the original signal with the fused signal. The *Transformer* model uses a transformer to combine these two signals. Across all metrics for both CMU-MOSEI and CH-SIMS, these three methods exhibit similar performance.

Context window	$\mathbf{Has0\_ACC_2}$	$\mathbf{Has0\_F1}$	$\mathbf{Non0\_ACC_2}$	$\mathbf{Non0\_F1}$	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	$\mathbf{MAE}$	$\mathbf{Corr}$
0	84.89	84.86	87.04	87.07	**54.81**	**47.32**	66.62	83.27
1	**86.01**	**85.94**	**88.01**	**87.99**	53.35	45.87	**66.37**	**83.96**
2	85.81	85.71	87.8	87.76	53.98	46.99	66.7	82.49
3	84.94	84.86	86.84	86.81	52.14	45.19	69.85	81.3

(a) Concatenation

Context window	$\mathbf{Has0\_ACC_2}$	$\mathbf{Has0\_F1}$	$\mathbf{Non0\_ACC_2}$	$\mathbf{Non0\_F1}$	$\mathbf{ACC_5}$	$\mathbf{ACC_7}$	$\mathbf{MAE}$	$\mathbf{Corr}$
0	84.89	84.86	87.04	87.07	54.81	47.32	66.62	83.27
1	85.57	85.51	87.8	87.8	**55.44**	**47.81**	65.17	83.37
2	**86.2**	**86.12**	**88.46**	**88.44**	55.24	47.04	**63.88**	**84.46**
3	85.76	85.69	88.01	87.99	54.67	46.89	65.26	83.71

(b) Separation

Table 7: **Comparison of Context Modeling Methods with Only Text signals**: the second method that separates the context and the current utterance can handle a longer context window and have a better performance.

	$\mathbf{Has0\_ACC_2}$	$\mathbf{Has0\_F1}$	$\mathbf{Non0\_ACC_2}$	$\mathbf{Non0\_F1}$	$\mathbf{ACC_7}$	$\mathbf{MAE}$	$\mathbf{Corr}$
fusion network	**85.91**	**85.85**	**88.16**	**88.15**	**48.25**	**64.29**	**83.8**
- self attention layers	85.13	85.12	87.35	87.38	46.94	65.81	81.53
- fully connected layers	85.13	85.09	87.2	87.2	46.65	66.48	83.07

Table 8: **Fusion Network Ablation Experiment Results**