 ISSN (Online): 3140-5312

Multi-scale Temporal-aware Text-Enhanced Network for Multimodal Sentiment Analysis

Junjun Zhang[1], Giseop Noh[2]
1.School of Software, Henan University of Engineering, Zhengzhou 451150, China
2. Department of Software and Communications Engineering, Hongik University, 2639 Sejong-ro, Jochiwon-eup, Sejong-si, 30016, Republic of Korea
Corresponding author: Giseop Noh (e-mail: kafa46@hongik.ac.kr)

Abstract

Multimodal Sentiment Analysis (MSA) has achieved remarkable progress by leveraging textual, visual, and acoustic cues. However, adequately modeling the dynamic, fine-grained temporal dependencies both within and across modalities in long video sequences remains a formidable challenge. Existing Transformer-based fusion networks, while adept at capturing cross-modal interactions, often fall short in comprehensively extracting and integrating sentiment evolutions across different temporal scales. Building upon the prevailing paradigm of text-enhanced Transformer fusion networks, this paper addresses their limitation in modeling fine-grained temporal dependencies in long video sequences by proposing a novel Multi-scale Temporal-aware Speech-Enhanced Transformer Fusion Network (MSTETFN). Specifically, we introduce: 1) a Multi-scale Temporal Sequence Learning Module (MTS-TLM) to capture sentiment variations at different temporal granularities; 2) a Hierarchical Temporal Attention Mechanism (HTAM) to dynamically weight and aggregate these multi-scale temporal features; and 3) a Temporally-Aligned Text-Enhanced Transformer (TA-TET) to integrate these temporally rich representations into cross-modal interactions. Extensive experiments conducted on the benchmark datasets CMU-MOSI and CMU-MOSEI demonstrate that MSTETFN significantly outperforms current state-of-the-art methods, particularly for long video segments and complex sentiment dynamics.

Keywords: Multimodal Sentiment Analysis; Temporal Modeling; Multi-scale Feature Learning; Hierarchical Attention

1. Introduction

The proliferation of multimedia content on social platforms has propelled Multimodal Sentiment Analysis (MSA) into a research hotspot. Unlike unimodal approaches, MSA integrates complementary cues from text, audio, and visual modalities to achieve a more comprehensive understanding of human affective states, a process that mimics the human cognitive mechanism of expressing emotions through language, facial expressions, and vocal intonation. Early MSA research primarily focused on fusion strategies, evolving from simple concatenation to decision-level integration [1], with subsequent advancements in attention mechanisms and tensor fusion networks [2–3] further enhancing performance. The advent of the Transformer architecture [4], particularly its self-attention and cross-attention mechanisms, has enabled Transformer-based MSA models to effectively model complex intra-modal and inter-modal relationships, yielding significant progress. Among various approaches, the text-enhanced Transformer fusion network has garnered considerable attention by recognizing the central role of text in emotional expression. Text typically conveys the most explicit sentiment information; such methods augment non-verbal modalities through text-oriented cross-modal mappings, while incorporating unimodal label prediction to preserve modality-specific characteristics. However, despite their notable fusion performance, these methods remain deficient in modeling fine-grained temporal dependencies and multi-scale sentiment evolution within long video sequences.

Human affect is inherently dynamic, evolving at varying speeds and intensities over time. Emotions may shift abruptly within seconds—such as a sudden burst of laughter—or gradually unfold over longer durations, as seen in sustained irony. Accurately capturing these different temporal granularities is crucial for sentiment analysis in real-world scenarios. Existing Transformer

 ISSN (Online): 3140-5312

models typically process sequences at a fixed temporal resolution or rely on generic temporal encoders, making it difficult to simultaneously capture instantaneous reactions and overarching affective trajectories. This limitation is particularly pronounced when facing subtle or delayed sentiment transitions, often resulting in an incomplete understanding of affective dynamics.

To address this critical issue, we propose a novel Multi-scale Temporal-aware Text-Enhanced Transformer Fusion Network (MSTETFN). This network directly embeds sophisticated multi-scale temporal awareness into the feature learning and fusion processes. Our main contributions are as follows:

(1) Multi-scale Temporal Sequence Learning Module (MTS-TLM). We introduce parallel temporal convolutional networks with varying kernel sizes and dilation rates. This module independently processes features from each modality to extract sentiment-relevant patterns at different temporal granularities.

(2) Hierarchical Temporal Attention Mechanism (HTAM). To intelligently integrate the diverse information derived from MTS-TLM, we design a hierarchical attention mechanism. HTAM dynamically assigns weights to features across different temporal scales for each modality, enabling the model to adaptively focus on the most relevant temporal resolutions according to the specific affective characteristics of the input.

(3) Temporally-Aligned Text-Enhanced Transformer (TA-TET). The refined multi-scale temporal-aware features are subsequently fed into a modified text-enhanced Transformer. This ensures that cross-modal interactions and text-guided augmentations are performed on representations that inherently preserve and exploit fine-grained temporal context, thereby facilitating more robust and temporally coherent multimodal fusion.

(4) We conduct comprehensive experiments and ablation studies on two benchmark datasets, demonstrating the state-of-the-art performance of MSTETFN and validating the effectiveness of its novel components.

2. Related Work

2.1 Multimodal Sentiment Analysis (MSA)

MSA aims to overcome the limitations of unimodal analysis by integrating information from multiple modalities. Early fusion concatenates features from different modalities before feeding them into a single model. Although simple, it often fails to capture complex cross-modal relationships and is sensitive to misalignment. Late fusion processes each modality independently and fuses their predictions at the decision level [1]. This approach preserves modality-specific information but may miss deep and subtle cross-modal interactions. Intermediate fusion combines modalities within the neural network architecture. Methods such as TFN [5] and LMF [6] leverage tensor operations to model outer-product interactions. Attention-mechanism-based fusion models [2] utilize cross-modal attention to facilitate information exchange across modalities. Some advanced methods [7–8] further refine fusion by treating text as the core modality and augmenting non-verbal representations during cross-modal mapping..

2.2 Temporal Modeling in MSA

The temporality of emotional expression is crucial in MSA, as many tasks involve analyzing video or audio segments where sentiments evolve dynamically. Early recurrent neural networks were widely used to capture sequential dependencies within each modality, and CHFusion [9] further employed a hierarchical RNN structure to model fine-grained local correlations. However, traditional RNNs struggle with very long sequences and implicitly capture temporal dynamics without explicitly considering multi-scale patterns. TCN[10] emerged as an alternative, offering parallelism and long-range dependency capture through dilated convolutions, yet they typically operate with a single fixed receptive field. Transformers, equipped with positional encoding and self-attention, inherently process sequence data and capture global temporal context, but they treat all time steps uniformly, which may be suboptimal for discerning sentiment variations across different temporal scales.

Our proposed MSTETFN uniquely combines the strengths of the text-enhanced fusion paradigm with an explicit, multi-scale, and hierarchically attentive temporal modeling strategy. Unlike prior methods that focus on single-scale dependencies or global context, MSTETFN ensures that fine-grained affective dynamics across various temporal resolutions are captured, dynamically weighted, and effectively integrated into cross-modal fusion. This direct remedy to the multi-scale temporal limitation constitutes a novel contribution to MSA.

3. Methodology

In this section, we elaborate the architecture of our proposed Multi-scale Temporal-aware Text-Enhanced Transformer Fusion Network (MSTETFN). We first formulate the problem and present the overall architecture and then describe each component in detail.

 ISSN (Online): 3140-5312

3.1 Problem Formulation and Overall Architecture

The task of Multimodal Sentiment Analysis (MSA) is to predict the sentiment score $y \in [-3,3]$ for a given video segment, which contains three modalities: text I_t, visual I_v, and acoustic I_a. Each modality $m \in \{t,v,a\}$ is represented as a sequence $I_m \in \mathbb{R}^{T_m \times d_m}$. As shown in Figure 1, MSTETFN consists of four main stages: (1) feature extraction and initial context encoding; (2) Multi-scale Temporal Sequence Learning Module (MTS-TLM); (3) Hierarchical Temporal Attention Mechanism (HTAM); (4) Temporally-Aligned Text-Enhanced Transformer (TA-TET) and Unimodal Label Generation Module (ULGM) for final sentiment prediction.

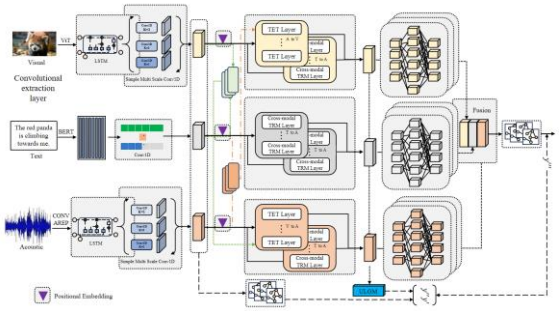

Fig. 1. Architecture diagram of the MSTETFN model

3.2 Feature Extraction and Initial Context Encoding

Drawing upon the successful paradigm of existing text-enhanced fusion networks, we leverage robust pre-trained models for unimodal feature extraction to guarantee high-fidelity initial representations. Specifically, we adopt BERT for textual encoding, ViT for visual encoding, and the COVAREP toolkit for acoustic encoding, yielding the initial feature sets F_t, F_v and F_a for the three respective modalities.

3.3 Multi-scale Temporal Sequence Learning Module (MTS-TLM)

MTS-TLM aims to explicitly capture sentiment-relevant patterns at different temporal granularities. For each modality m, we adopt a parallel architecture of Temporal Convolutional Networks (TCN), specifically one-dimensional convolutional layers with varying kernel sizes and optional dilation rates. Each branch learns features at a distinct temporal scale.

Let h_m^0 be the input sequence for modality m. For each modality, we define k parallel convolutional branches. Each branch $k \in \{1,2,\cdots,K\}$ applies a one-dimensional convolution with a specific kernel size s_k and dilation rate d_k:

$$F_m^{scale,k} = Conv1D(h_m^0; Kerel\_size = s_k, dilation\_rate = d_k) \quad (1)$$

where $F_m^{scale,k} \in \mathbb{R}^{T_m \times d^t}$ denotes the features extracted at scale k. Small kernel sizes with a dilation rate of 1 capture fine-grained, immediate temporal dynamics, while large kernel sizes or dilated convolutions capture broader contextual temporal patterns. The outputs of the parallel branches are then concatenated:

$$F_m^{multi-scale} = ConcatD(F_m^{scale,1}, F_m^{scale,1}, \cdots, F_m^{scale,K}) \quad (2)$$

3.4 Hierarchical Temporal Attention Mechanism (HTAM)

The MTS-TLM provides a rich set of multi-scale temporal features. The Hierarchical Temporal Attention Mechanism (HTAM) is designed to dynamically weight these multi-scale features. For each modality m, given its multi-scale feature representation $F_m^{multi-scale} \in \mathbb{R}^{T_m \times (K \cdot d')}$, we first split it back into its K scale-specific components. Then, HTAM computes attention weights over these K scales:

$$a_{m,t,k} = \frac{exp(W_{HTAM} \cdot F_{m,t}^{scale,k})}{\sum_{j=1}^{K} exp(W_{HTAM} \cdot F_{m,t}^{scale,k})} \quad (3)$$

where $a_{m,t,k}$ is the attention weight for modality $t$at time step k and scale t. The final temporal-aware feature representation $H_m \in \mathbb{R}^{T_m \times d'}$ for modality m is obtained by a weighted sum over the scale-specific features.

$$H_{m,t} = \sum_{k=1}^{K} a_{m,t,k} \cdot F_{m,t}^{scale,k} \quad (4)$$

3.5 Temporally-Aligned Text-Enhanced Transformer (TA-TET)

The core fusion mechanism of MSTETFN draws upon the idea of the text-enhanced Transformer. Let

 ISSN (Online): 3140-5312

H_t, H_v, H_a be the temporal-aware representations from HTAM. First, to augment the non-verbal modalities with textual information, H_v and H_a query H_t.

$$c_v = Mul\_Att(Q = H_v W_{Q_v}, K = H_t W_{K_{t1}}, V = H_t W_{V_{t1}}) \tag{5}$$

$$c_a = Mul\_Att(Q = H_a W_{Q_a}, K = H_t W_{K_{t2}}, V = H_t W_{V_{t2}}) \tag{6}$$

Here, c_v and c_a are intermediate representations of the visual and acoustic modalities.

Next, we model pairwise cross-modal mappings. For the i-th layer:

$$h^i_{a\to v} = Mul\_Att(Q = H_v W'_Q, K = c_a W'_{K_a}, V = c_a W'_{V_a}) + LN(H_v) \tag{7}$$

$$h^i_{t\to v} = Mul\_Att(Q = H_v W''_{Q_v}, K = H_t W''_{K_t}, V = H_t W''_{V_t}) + LN(H_v) \tag{8}$$

After L layers, the final text-enhanced representation of the visual modality becomes:

$$H_v^{fused} = Concat(h^L_{a\to v}, h^L_{t\to a}) \tag{9}$$

Similarly, we obtain H_v^{fused} by concatenating $h^L_{a\to v}$ and $h^L_{t\to a}$. For the text modality itself, we apply self-attention to H_t to gather its intrinsic temporal context:

$$H_t^{fused} = s\_Att(H_t) + LN(H_t) \tag{10}$$

3.6 Unimodal Label Generation Module (ULGM)

For each modality $m \in \{t, v, a\}$, its final temporal-aware representation H_m is passed through a modality-specific fully connected layer to predict a unimodal sentiment score.

$$\hat{y}_s^m = FC_m(H_m^{pooled}) \tag{11}$$

where H_m^{pooled} is typically the output at the final time step or an average pooling of H_m.

3.7 Sentiment Prediction and Loss Function

The final fused representations H_t^{fused}, H_v^{fused}, and H_a^{fused} are concatenated and passed through a fully connected layer to generate the multimodal sentiment prediction:

$$\hat{y}_{mul} = FC_{final}(Cpncat(H_t^{fused}, H_v^{fused}, H_a^{fused})) \tag{12}$$

The total loss function is a combination of the multimodal regression loss and the unimodal prediction loss.

$$\mathcal{L}_{total} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}^i_{mul} - y^i| + \lambda \frac{1}{N}\sum_{m\in\{t,a,v\}}^{N} w_i^m \cdot |y_s^{m,(i)}| \tag{13}$$

where λ is a hyperparameter balancing the multimodal and unimodal tasks.

4. Experiments

4.1 Datasets

We evaluate MSTETFN on two widely used benchmark datasets for multimodal sentiment analysis. The CMU-MOSI dataset contains short video clips from YouTube reviews, each manually annotated with sentiment intensity labels ranging from -3 to +3. CMU-MOSEI is the largest MSA dataset, containing 23,353 annotated video clips from 2,928 videos.

4.2 Evaluation Metrics

We adopt standard metrics for regression and classification tasks. For regression, we employ Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr). For classification, we binarize sentiment scores (negative vs. non-negative / negative vs. positive). We report binary classification accuracy (Acc-2), i.e., the percentage of correctly classified samples, for both "negative vs. non-negative" and "negative vs. positive" classifications.

4.3 Baseline

To comprehensively evaluate the performance of MSTETFN, we compare it with several state-of-the-art MSA models:

- TNF [3]: Tensor Fusion Network, an early and influential intermediate fusion method.
- LMF [6]: Low-rank Multimodal Fusion, an efficient variant of tensor fusion.
- MulT [2]: Multimodal Transformer, which employs cross-modal Transformers for modality conversion.
- MISA[7]: Modality-Invariant and Modality-Specific Representations, which learns disentangled features.

TETFN [11]: Text-Enhanced Transformer Fusion Network.

 ISSN (Online): 3140-5312

4.4 Implementation Details

The experiments employ the Adam optimizer, with a fine-tuning learning rate of 5×10^{-5} for BERT and 1×10^{-3} for other parameters, and a batch size of 64. The LSTM hidden size is 128. The Conv1D kernel size for all modalities is 3, and the visual feature length is 50 frames. For the Multi-scale Temporal Sequence Learning Module (MTS-TLM), we employ K = 3 parallel one-dimensional convolutional branches for each modality. Scale 1 uses a kernel size of 3 with a dilation rate of 1; Scale 2 uses a kernel size of 5 with a dilation rate of 1; Scale 3 uses a kernel size of 3 with a dilation rate of 2.

HTAM module uses 5 heads for the text-oriented multi-head attention in the Transformer layers. The cross-modal Transformer employs L = 3 layers. The loss weight λ is empirically set to 0.7 for MOSI and 0.5 for MOSEI. All models are trained for 50–100 epochs with early stopping based on validation MAE. We report average results over 5 runs with different random seeds.

4.5 Quantitative Results

Table 1. presents the performance of MSTETFN and various baseline models on the CMU-MOSI dataset. All results are averaged over multiple runs

Model	MAE↓	Corr↑	Acc-2	F1-score
LMF(B)	0.917	0.695	-/82.5	-/82.4
MulT(B)	0.861	0.711	81.5/84.1	80.6/83.9
MISA(B)	0.783	0.761	81.8/83.4	81.7/83.6
TETFN*	0.717	0.800	84.05/86.10	83.83/86.07
MSTETFN (Ours)	**0.693**	**0.812**	**85.31/87.19**	**85.15/87.19**

Table 2. Performance results on the CMU-MOSEI dataset

Model	Corr↑	Acc-2	F1-score
TNF(B)	0.700	-/82.5	-/82.1
LMF(B)	0.677	-/82.0	-/82.1
MulT(B)	0.703	-/82.5	-/82.3
MISA(B)	0.756	83.6/85.5	83.3/85.3
TETFN*	0.748	84.25/85.18	84.18/85.27
MSTETFN (Ours)	**0.771**	**85.46/86.32**	**85.39/86.28**

In terms of regression, MSTETFN achieves an MAE of 0.693 and a Pearson correlation of 0.812 on CMU-MOSI, and 0.522 and 0.771 on CMU-MOSEI, outperforming all baselines including TETFN. The higher correlation indicates more accurate sentiment intensity capture. For classification, MSTETFN attains Acc-2 of 85.31/87.19% on CMU-MOSI and 85.46/86.32% on CMU-MOSEI, demonstrating enhanced polarity identification even in complex cases. The consistent improvements, particularly on the larger and more diverse CMU-MOSEI dataset, strongly suggest that explicitly modeling and dynamically weighing multi-scale temporal dependencies significantly enhances the network's understanding of affective dynamics.

4. Conclusion

In this paper, we have proposed a Multi-scale Temporal-aware Text-Enhanced Transformer Fusion Network (MSTETFN), aiming to overcome the limitation of existing multimodal sentiment analysis models in modeling fine-grained temporal dependencies in long video sequences. Building upon the solid foundation established by the text-enhanced Transformer fusion network paradigm, MSTETFN integrates three key innovations: a Multi-scale Temporal Sequence Learning Module (MTS-TLM) for extracting sentiment-relevant patterns at different temporal granularities, a Hierarchical Temporal Attention Mechanism (HTAM) for dynamically weighting and aggregating these multi-scale features, and a Temporally-Aligned Text-Enhanced Transformer (TA-TET) for robust cross-modal interactions.

Our comprehensive experimental evaluations on two publicly available benchmark datasets demonstrate that MSTETFN achieves state-of-the-art performance, validating the effectiveness of our proposed multi-scale temporal awareness and hierarchical attention mechanism..

Conflict of Interest

The authors declare that there are no potential conflicts of interest related to this paper.

Acknowledgments

 ISSN (Online): 3140-5312

This work was supported by the Henan Provincial Key Scientific Research Projects Program and the for Higher Education Institutions under Grant 25A520053 and National Innovation and Entrepreneurship Training Program for College Students 2026 (Grant No. S202611517033).

References

[1] Poria, S., Cambria, E., Howard, N., Huang, G. B., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, pp. 98-125.

[2] Tsai, Y.-H.H.,Bai, S.,Liang, P.P.,Kolter, J.Z., Morency, L.-P. Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019, pp. 6558–6569..

[3] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency.Tensor fusion network for multimodal sentiment analysis. In Empirical Methods in Natural Language Processing, EMNLP.Copenhagen, Denmark,Sep. 9-11, 2017, pp. 1103–1114, doi:10.18653/v1/D17-1115.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems, 30, pp. 5998-6008, 2017.

[5] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI conference on artificial intelligence ,Feb. 2–7, 2018, 32(1),pp.5622–5630.

[6] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,Melbourne, Australia,2018, vol.1, pp.2247–2256.

[7] Wang, D., Guo, X., Tian, Y., Liu, J., He, L., & Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition, vol. 136, pp. 109259, 2023.

[8] Hou, J., Omar, N., Tiun, S., Saad, S., & He, Q. . TCHFN: Multimodal sentiment analysis based on text-centric hierarchical fusion network. Knowledge-Based Systems, vol.300, pp.112220 , 2024.

[9] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) ,pp. 1724–1734 , 2014.

[10] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. Temporal Convolutional Networks for Action Segmentation and Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , pp. 1003-1012 , 2017.

[11] Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131. doi: 10.1145/3394171.3413678.

Author Information

Junjun Zhang
2008: BE, Computer Science and Technology, Nanyang Institute of Technology, China
2017: ME, Computer Technology, Zhengzhou University, China
2023: PhD, Computer Information Engineering, Cheongju University, South Korea
2009~2019: Professor, School of Software, Zhengzhou University of Industrial Technology,China
2024~Present: Professor, School of Software, Henan University of Engineering,China
Email: zhangjunjun@zzuit.edu.cn
Research Interest: AI, NLP, Computing, Emotion Analysis

Giseop Noh
1998: BE, Industrial Engineering, Korea Air Force Academy, Korea
2009: MS, Computer Science, University of Colorado Denver, USA
2014: PhD, Computer Science, Seoul National University, Korea
1994~2005: Officer, Republic of Korea Air Force, responsible for operation and maintenance of advanced weapon computer systems
2017~2018: Faculty, Department of Computer Science, Korea Air Force Academy
2018~2024: Professor, Department of Artificial Intelligence Software, Cheongju University, Korea
2024~Present: Professor, Department of Software and Communications Engineering, Hongik University
Email: kafa46@hongik.ac.kr
Research Interest: AI, NLP, Computing