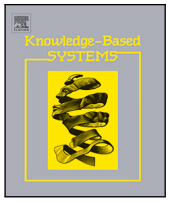
Knowledge-Based Systems 300 (2024) 112220

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



TCHFN: Multimodal sentiment analysis based on Text-Centric Hierarchical Fusion Network

Jingming Hou [a], Nazlia Omar [a,*], Sabrina Tiun [a], Saidah Saad [a], Qian He [b]

[a] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43650, Selangor, Malaysia
[b] State and Local Joint Engineering Research Center for Satellite Navigation and Location Service, Guangxi Key Laboratory of Cryptography and Information Security, Guangxi Collaborative Innovation Center of Cloud Computing and Big Data, Guilin University of Electronic Technology, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Text-centric
Fusion network
Transformer
Contrastive learning
Knowledge distillation

ABSTRACT

Multimodal sentiment analysis (MSA) has become a popular field of research in recent years. The aim is to combine the three modalities of text, video, and audio to obtain comprehensive emotional information. However, current research often treats these three modalities equally, downplaying the crucial role of text modality in MSA and ignoring the redundant information generated during multimodal fusion. To address these problems, we propose the Text-Centric Hierarchical Fusion Network (TCHFN), employing a hierarchical fusion strategy. In this framework, low-level fusion involves cross-modal interactions between pairs of modalities, while high-level fusion extends these interactions to involve all three modalities. Through the design of the Cross-modal Reinforced Transformer (CRT), we achieve cross-modal enhancement of the target modality, facilitating a nuanced fusion process with text serving as its core. Additionally, we design Text-Centric Contrastive Learning (TCCL) to align non-text modalities with the text modality, emphasising the central role of text in the fusion process. After fusion, a multimodal fusion output gate is employed to mitigate redundant information within the multimodal fusion representation, which is subsequently processed by a linear layer for prediction. Simultaneously, to fully leverage limited labelled datasets, we introduced knowledge distillation. This approach involves preserving the model parameters that yield the best performance during training as a teacher model. The teacher model aids in capturing rich emotional information, enabling the model to transcend local optima and discover more optimal parameters, thereby enhancing overall model performance. Extensive experiments on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate the superiority of our model over state-of-the-art methods in MSA tasks.

1. Introduction

With the advancement of technology, the rise of multimedia applications such as TikTok and Instagram has generated a vast amounts of multimodal data, including text, acoustics, and visuals [1]. Customers can easily document their lives using smartphones and upload these short videos on various platforms. However, the abundance of videos poses new challenges for sentiment analysis [2]. This requires machines to effectively understand sentiment information embedded in multimodal data. In response to this, multimodal sentiment analysis (MSA) has emerged, aiming to amalgamate and harness diverse information sources [3]. This integration provides a more holistic perspective for enhancing the depth of sentimental understanding. As shown in Fig. 1, a happy tone and smiling face can assist in more accurately predicting positive sentiments in emotionally ambiguous sentences. Therefore, MSA has extensive practical applications. For instance, chatbots can utilise sensors to capture multimodal data and tailor subsequent responses based on emotional predictions from the MSA model, thereby creating more engaging and contextually adaptive interactions.

In the early stages, to obtain effective multimodal representations, it was common to directly concatenate various modalities and then fuse them through various transformations using Long Short-Term Memory (LSTM) [4–6]. However, this fusion approach failed to effectively capture the correlations between emotional features across different modalities. As Transformer have demonstrated state-of-the-art performance in various domains, the use of Transformer architectures in the multimodal field has gradually increased. The parallel mechanism of the Transformer boosts the efficiency of the MSA model, while the attentional mechanism of the Transformer adeptly captures the emotional correlations across modalities. Rahman et al. [7] introduced a Multimodal Adaptation Gate in Bidirectional Encoder Representations

* Corresponding author.
E-mail addresses: houjimmy1@163.com (J. Hou), nazlia@ukm.edu.my (N. Omar), sabrinatiun@ukm.edu.my (S. Tiun), saidah@ukm.edu.my (S. Saad), heqian@guet.edu.cn (Q. He).

https://doi.org/10.1016/j.knosys.2024.112220
Received 27 January 2024; Received in revised form 14 June 2024; Accepted 7 July 2024
Available online 14 July 2024

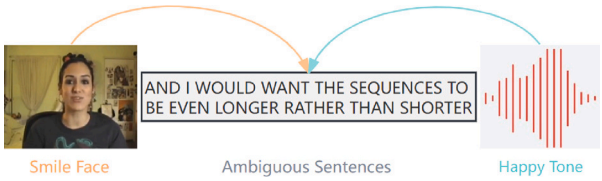


Fig. 1. Example of non-text modality assisting text modality on CMU-MOSI dataset.

from Transformers (BERT) to enable its internal components to accept non-text modalities as inputs. To establish the remote cross-modal interactions of raw features, Tsai et al. [8] proposed the Multimodal Transformer (MulT) model, making the Transformer more suitable for multimodal tasks. Subsequently, He et al. [9] proposed the Unimodal Reinforced Transformer, inspired by the MulT model. Currently, numerous experiments are being conducted to explore similar and dissimilar information among different modalities, with the aim of reducing the gap and heterogeneity between different modalities. Hazarika et al. [10] presented the MISA framework, projecting each modality into two different subspaces to identify commonalities and unique information in multimodal data. Yang et al. [11] proposed the ConFEDE framework, which decomposes the representation of each modality and uses contrastive learning to distinguish similar and dissimilar features between modalities.

Despite the promising results achieved by existing methods, they mainly treat each modality equally without distinguishing their importance in sentiment prediction. For MSA, the text modality tends to perform the best for several reasons: (1) Non-text (acoustic and visual) data often incorporate more noise than text, owing to environmental and equipment-related factors [12]. (2) Sampling methods for non-text modalities may miss crucial segments [13], while text directly vectorises each word, capturing more comprehensive information. (3) Text inherently carries more emotional information than non-text data [14]. (4) Pre-trained models such as BERT and Generative Pre-trained Transformer (GPT) significantly boost the capabilities of text sentiment analysis [15]. Recent studies have attempted to enhance the text modality with non-text modalities [16,17]; however, they still have not fully leveraged the inherent advantages of the text modality. Moreover, after multimodal fusion, the resulting representations often include redundant information [18], which typically influences prediction outcomes. Previous research has predominantly ignored this issue by directly inputting multimodal representations into linear networks for classification.

To fully exploit the core advantages of the text modality in MSA and obtain more effective multimodal fusion representations, we introduce the Text-Centric Hierarchical Fusion Network (TCHFN). This network comprises low-level and high-level fusions. The low-level fusion focuses on establishing cross-modal interactions between text and non-text modalities and enhancing representations of text, acoustic and visual data. The high-level fusion emphasises text centric cross-fusion by enhancing text-acoustic and text-visual fusion representations, thereby resulting in more effective tri-modal fusion representations. Inspired by MulT [8], we design the Cross-modal Reinforced Transformer (CRT) to achieve the cross-modal enhancement of the target modality during the fusion process. In low-level fusion, non-textual information is dynamically added to textual representations by combining CRT with a dynamic attention mechanism, and then concatenating the utterance-level features of the text to form a Text Enhancement Network. Using non-textual representations as the target modality of CRT, we enable non-textual representations to capture sentiment information from the text. Additionally, we design Text-Centric Contrastive Learning (TCCL) to further enhance non-text representations by aligning them with text representations forming a Non-text Enhancement Network. In high-level fusion, CRT is used for cross-fusion to obtain a trimodal fusion representation with emotionally rich representations, while the Multimodal Representation Output Gate module, inspired by LSTM gates is used to reduce redundancy in the trimodal fusion representation.

Given the limited labelled data for MSA [19], we introduce knowledge distillation to maximise the available data. This entails preserving the parameters of the best-performing model as a teacher model during training to assist subsequent training phases, thereby enhancing the model's performance. To the best of our knowledge, this is the first study to introduce knowledge distillation into the field of MSA.

Finally, we evaluate our model by performing sentiment prediction tasks on three benchmark datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. Experimental results demonstrate that our model outperforms state-of-the-art methods across various metrics. Furthermore, ablation studies are also conducted to confirm the effectiveness of our model.

In summary, our contributions include:

(1) Proposing TCHFN to fully leverage the core advantages of the text modality in MSA. This network involves designing a CRT for cross-modal interaction and target modality enhancement, as well as introducing TCCL to align non-text modalities with the text modality.

(2) Designing a Multimodal Representation Output Gate to automatically select and output useful information from multimodal representations, thereby reducing redundancy in the fusion process.

(3) Introducing knowledge distillation for the first time in the field of MSA, fully leveraging existing data for training to enhance the model's performance.

(4) Achieving superior results on different MSA benchmarks compared with existing state-of-the-art methods.

The remainder of this paper is organised as follows: Section 2 reviews related works on MSA. Section 3 provides a detailed introduction to TCHFN. Section 4 introduces the experimental setup and the datasets used. Section 5 presents the analysis and comparison of the experimental results. Finally, Section 6 concludes the paper.

2. Related works

2.1. Unimodal sentiment analysis

Unimodal sentiment analysis is a highly popular research field, with significant research value in sentiment analysis for text, acoustic, and visual modalities [20,21]. Sentiment analysis of text has been a mainstream focus [22], with early studies relying on sentiment dictionaries [23,24]. However, this method requires manual updating of dictionaries, which in turn necessitates a considerable amount of human resources [25,26]. With technological advancements, machine learning, including algorithms such as Support Vector Machines (SVM) [27], Decision Trees [28], Conditional Random Fields (CRF) [29] have gained popularity because they do not require sentiment dictionaries and often achieve better predictive results with improved generalisation. Currently, the predominant approach is the use of deep learning, particularly large language models such as BERT [30] and GPT-3 [31], to enhance predictive accuracy by improving text feature representations. Extensive research has validated the effectiveness of this approach.

Secondly, sentiment analysis based on speech primarily focuses on studying the speaker's tone, pitch, and speed [32]. Computers can perform subsequent analyses and predictions by segmenting speech and converting the spectrograms of the speech segments into vector representations. However, differences in speech quality often result in significant noise, leading to prediction instability. Current methods predominantly involve Recurrent Neural Networks (RNN) [33] and Transformer [34].

Finally, sentiment analysis based on visual cues primarily involves studying a speaker's facial expressions, head movements, and changes in expressions [35]. Currently, Convolutional Neural Networks (CNN) have been proven effective in the field of image processing and have widespread applications [36]. With the success of transformers in natural language processing, computer vision has attracted researchers to

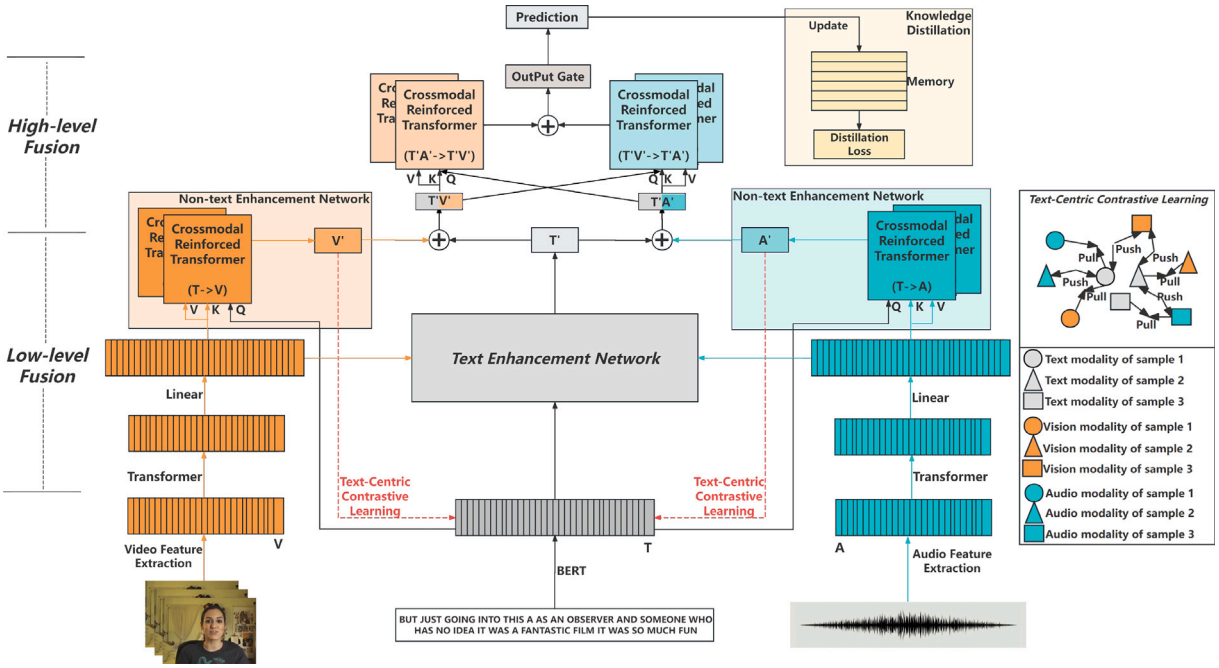


Fig. 2. Overall architecture of the TCHFN.

migrate Transformer to the field. Zhang et al. [37] proposed a feature pyramid transformer utilising self-attention, top-down cross-attention, and bottom-up cross-channel attention to model interactions across space and scales, achieving performance improvements across multiple tasks.

2.2. Multimodal sentiment analysis

MSA is crucial for understanding human emotions. Limiting the analysis to a single modality would overlook the valuable information conveyed through multiple modalities. Multimodal fusion technology is a critical focus in MSA research, aiming to effectively integrate heterogeneous information such as text, audio, and video to obtain more effective multimodal fusion feature representations, thus achieving significant progress [38].

Combining previous research, MSA methods can be categorised into four main approaches: early fusion-based methods, late fusion-based methods, feature space-based fusion methods, and translation-based fusion methods.

Early fusion methods typically involve extracting features from different modalities and then combining them through operations such as concatenation, addition, and multiplication to form a larger and more comprehensive feature representation [39]. For instance, Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) adopts early fusion, directly concatenating feature representations from different modalities and using them as inputs to a single LSTM for predicting sentiment labels [40].

Late fusion occurs after each modality has undergone high-level feature extraction and modelling. Zadeh et al. [41] designed the Multi-attention Recurrent Network (MARN) model by dynamically modelling single modalities using Long-short Term Hybrid Memory (LSTHM) and using Multi-attention Block (MAB) to accept hidden states from all LSTHMs at each time step to achieve late fusion through majority voting. Zeng et al. [42] proposed the Ensemble-based Missing Modality Reconstruction (EMMR) network for making better decisions through ensemble learning.

Feature space-based fusion methods involve mapping each modality to its respective feature space and learning the relationships between features through mathematical analysis or algorithms to enhance multimodal fusion effectiveness. Wang et al. [43] introduced the Recurrent Attended Variation Embedding Network (RAVEN), adding non-text modalities as biases to the corresponding text modalities. Hu et al. [44] introduced the Unified MSA and ERC (UniMSE) to capture the differences and consistencies between sentiments and emotions, using contrastive learning to map samples to the feature space. To identify specific emotions across modalities, Yu et al. [45] proposed Contrastive Knowledge Injection (ConKI) for obtaining specific knowledge representations through knowledge injection and distinguishing general knowledge representations from specific knowledge representations in space through contrastive learning. Mai et al. [19] introduced HyCon to learn intramodal, intermodal, and interclass information through mixed contrastive learning.

Translation-based fusion methods leverage translation to convert one modality into another, thereby establishing interactions between modalities. This approach draws inspiration from the success of Seq2Seq models in machine translation. Alternatively, this can be achieved by adjusting the encoder structure of the Transformer to better capture the interactions between words. Tang et al. [46] proposed the Coupled-Translation Fusion Network (CTFN), a model that facilitates two-modal cross-modal translation for reconstructing missing information in modalities. Wu et al. [47] proposed the Text-Centred Shared-Private framework (TCSP), which achieves more accurate sentiment prediction by identifying shared and private semantics between non-text and text modalities. Kim et al. [48] introduced the All-modalities-inOne BERT (AOBERT) as a pre-training model in the multimodal domain. Lv et al. [49] improved the MulT model and proposed the Progressive Modality Reinforcement (PMR) model for three-modal cross-modal fusion.

Although these approaches have yielded promising results, they predominantly treat the three modalities impartially, overlooking the significant role of the text modality in MSA. Consequently, we modify the structure of the transformer encoder to facilitate text-centric multimodal fusion. We employ TCCL to achieve spatial alignment from non-textual modalities to text, thus fully leveraging the potential of the text modality.

2.3. Knowledge distillation

Knowledge distillation is a transfer learning technique that enhances the performance of a relatively simple model by transferring knowledge from a complex model. This concept, which was first introduced by Hinton et al. [50], utilised soft labels from a teacher model to successfully transfer knowledge to a student model. Park et al. [51] introduced a novel approach known as relational knowledge distillation, which focuses on transferring the mutual relations between data examples. Additionally, Liu et al. [52] investigated the application of knowledge distillation for training compact semantic segmentation networks, specifically addressing the challenges in distilling knowledge from cumbersome networks. Inspired by previous research, we introduce knowledge distillation into MSA to effectively leverage limited MSA datasets and improve model performance.

3. Methodology

In this section, we provide an overview of the TCHFN. Subsequently, we provide a detailed explanation of each component within the TCHFN, including unimodal feature extraction, CRT, text enhancement network, non-text enhancement network, multimodal representation output gate, and knowledge distillation.

3.1. Overall architecture

As depicted in Fig. 2, TCHFN encompasses two main components: low- and high-level fusion. Low-level fusion, which is centred on text modality representation, enhances and aligns text, acoustic, and visual representations. High-level fusion combines the enhanced representations of the three modalities with text as the central modality to extract high-level features and obtain more effective multimodal fusion representations. Specifically, low-level fusion comprises two network models: text-enhanced and non-text-enhanced networks. The text enhancement network enhances the text representation in two ways. Firstly, it uses the CRT and a dynamic attention mechanism to dynamically add non-text modality emotional information to the text modality based on importance. This enables the text representation to achieve cross-modal enhancement. Secondly, it concatenates utterance-level enhancements containing information regarding long-term dependencies in the text, resulting in a final enhanced text representation. The main aim of a non-text enhancement network is to enhance non-textual representations using the text modality. It first employs CRT to enable non-text modalities to learn from the text modality, retaining information that is more relevant to the text modality. Subsequently, through TCCL, non-text modalities are aligned with the text modality, reducing the distance between the non-text and text modality spaces after CRT processing. This further combines the semantic representation of non-text modalities with text modalities, enhancing the core role of text in the fusion process and resulting in a final enhanced non-text representation. High-level fusion involves concatenating the enhanced text-video representation and the enhanced text-audio representation and using CRT for cross-fusion. This process captures high-level features among the three modalities, thereby forming a more informative multimodal fusion representation. Subsequently, unnecessary redundancies are eliminated through the multimodal representation output gate, resulting in a more effective multimodal fusion representation. Finally, the multimodal fusion representation is input into multiple linear layers for sentiment prediction, and the parameters of the best-performing model are saved as teacher model parameters to assist future model training.

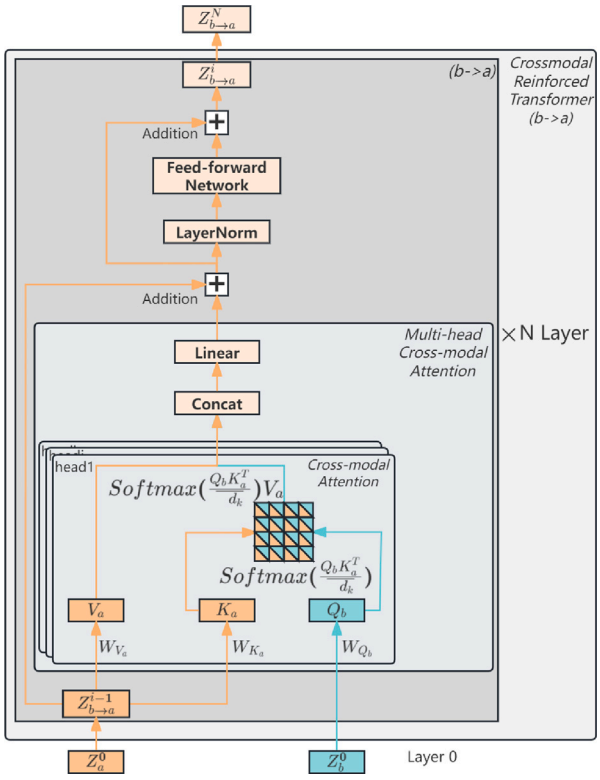


Fig. 3. CRT. Here, a represents the target modality, and b represents the source modality.

3.2. Unimodal feature extraction

We perform feature extraction and encoding using three pretrained models for each corresponding modality. The text modality uses the original sentence as input, aligning the sentence length with the other two modalities. If the sentence length exceeds L, truncation is applied; if it is less than L, padding is used. BERT is then employed for word embedding, resulting in a final feature representation of $T \in \mathbb{R}^{L\times d_t}$. For the audio and vision modalities, following a previous study [13,40], we employ pre-trained toolkits to extract the initial features. The resulting representation of the visual modality is denoted as $V \in \mathbb{R}^{L\times d_v}$, while the representation of the audio modality is denoted as $A \in \mathbb{R}^{L\times d_a}$, where L refers to the feature length. To capture the long-term dependencies between frames in video and audio and to enhance the understanding of modalities, the Transformer is applied to the feature representations of video and audio. Additionally, linear layers are used to align the dimensions of visual and acoustic feature representations with text modality feature representations, mapping $V \in \mathbb{R}^{L\times d_v}$ to $V \in \mathbb{R}^{L\times d_t}$ and $A \in \mathbb{R}^{L\times d_a}$ to $A \in \mathbb{R}^{L\times d_t}$. The processed text, visual, and acoustic representations are then used as inputs for the model.

3.3. Cross-modal reinforced transformer

Inspired by the MulT model [8] and based on the principles of the multi-head self-attention mechanism, we propose a CRT. We enhance the target modality by adjusting the cross-modal transformer in MulT. Specifically, the CRT linearly transforms the feature representation of the target modality into K and V, and the feature representation of the source modality into Q, where $K = Z_a^0 \times W_{K_a}$, $V = Z_a^0 \times W_{V_a}$ and $Q = Z_b^0 \times W_{Q_b}$. The remaining operations are identical to the cross-modal transformer in MulT. The process of computing multi-head cross-modal attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{1}$$

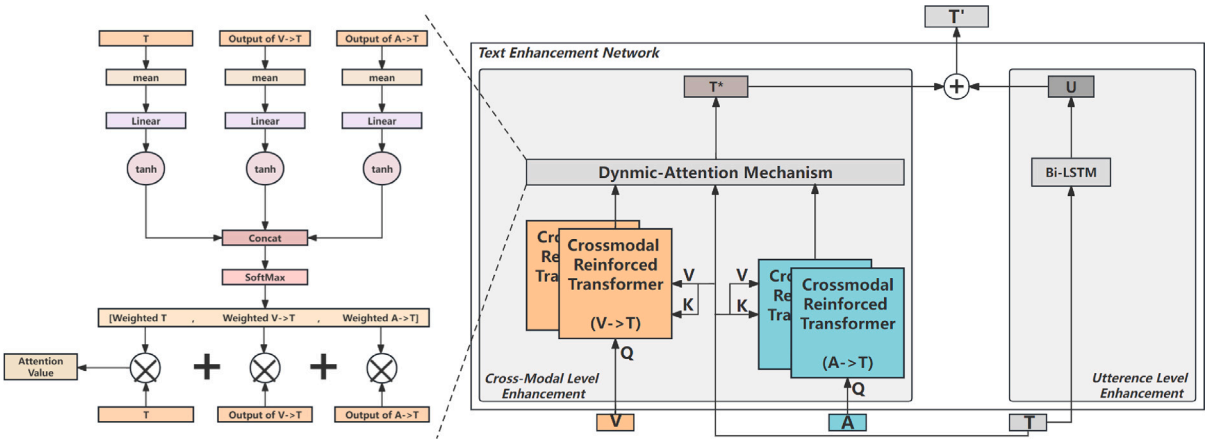


Fig. 4. Text enhancement network in TCHFN.

$$\text{head}_i = \text{Attention}(Q, K, V) \tag{2}$$

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \cdot W^O \tag{3}$$

where h is the number of heads in multi-head attention.

Subsequently, the output of multi-head attention undergoes processing through the Feed-forward network and LayerNorm network and is then input to the next layer for the same treatment. Crucial information related to the source modality of a target modality can be extracted through CRT. This facilitates the interactive exchange of information between the target and source modalities, thereby enhancing the feature representation of the target modality. The CRT is illustrated in Fig. 3.

3.4. Text enhancement network

The text enhancement network enhances text representations in two ways: cross-modal enhancement and utterance-level enhancement. In certain instances, the text may lack explicit emotional keywords, while non-text modalities (such as audio and visual cues) exhibit clear emotional indicators. For example, predicting the sentiment of the sentence 'Next week, I have a performance' based solely on its textual meaning can be challenging. However, integrating features such as the speaker's elevated voice and smiling expression enables the prediction of the speaker's positive emotional state. To achieve this, we use the CRT to make the text modality aware of important information from non-text modalities. The Dynamic Attention Mechanism enables the model to autonomously assess the importance of non-text modality emotional information for the text modality, dynamically incorporating it into the text modality. This results in cross-modal text enhancement. Simultaneously, since we adopt word-level representations, each word has unique feature representations. Despite the inter-word relationships acquired from the pre-trained BERT model, there is room for improvement in capturing context relationships for the overall sentence semantics. Thus, we input text feature representations into bidirectional LSTM (Bi-LSTM) to extract utterance-level representations. Finally, by concatenating the cross-modal enhanced representation T* with the utterance-level enhanced representation U, we obtain the final enhanced text representation T′. Specifically, text representation serves as the target modality, while acoustic and visual representations function as the source modalities for the CRT (see Eqs. (4) and (5)). Consequently, the text representation undergoes linear transformations to yield keys (K) and values (V), while the acoustic and visual representations are transformed into queries (Q). Subsequently, the text representation and two text representations processed by the CRT are fed into a dynamic attention mechanism for fusion. This process involves mean pooling, linear layers, and a tanh activation function. Following this, a weighted sum is employed to compute the cross-modal enhancement $T^* \in \mathbb{R}^{L \times d_t}$ (see Eq. (6)). Finally, due to the utilisation of Bi-LSTM for sentence-level enhancement, the dimensions of vector $U \in \mathbb{R}^{L \times 2}$ (see Eq. (7)). By concatenating U with T*, the ultimate text-enhanced representation $T' \in \mathbb{R}^{L \times (d_t + 2)}$ is obtained (see Eq. (8)). The Text Enhancement Network is illustrated in Fig. 4.

$$V \to T = CRT(V, T) \tag{4}$$

$$A \to T = CRT(A, T) \tag{5}$$

$$T^* = Dyn - Att(T, V \to T, A \to T) \tag{6}$$

$$U = Bi - Lstm(T) \tag{7}$$

$$T' = Concat(T^*; U) \tag{8}$$

3.5. Non-text enhancement network

The non-text modality includes acoustic and visual modalities, and both modalities undergo enhancement in the same manner. Firstly, the CRT enables non-text modalities to acquire information from the text modality. The text modality is linearly transformed into Q, and the non-text modality is linearly transformed into K and V. Through CRT, enhanced visual and acoustic feature representations $V' \in \mathbb{R}^{L \times d_t}$ (see Eq. (9)) and $A' \in \mathbb{R}^{L \times d_t}$ (see Eq. (10)) are obtained.

$$V' = T \to V = CRT(T, V) \tag{9}$$

$$A' = T \to A = CRT(T, A) \tag{10}$$

Subsequently, to further enhance and align the non-text modality with the text modality, we design Text-Centric Contrastive Learning (TCCL) to minimise the distance between the non-text and text modality spaces. Consequently, V′ and A′ are contrastively learned with T to reduce the distance between thw non-text and text modality spaces, further aligning the three modalities. The TCCL is illustrated in Fig. 2. Specifically, in a batch with n samples, positive samples consist of pairs of audio- and video-texts from the same sample (2n pairs), and negative samples consist of pairs of audio- and video-texts from different samples ($2n^2$-2n pairs). Contrastive learning aims to pull positive samples closer while pushing negative samples apart. The grey graphics represent text modality samples, orange-yellow graphics represent visual modality samples, and green graphics represent acoustic modality samples. Different colours with the same shapes indicate different modalities in the same sample. Contrastive learning loss is calculated using the Euclidean distance formula.

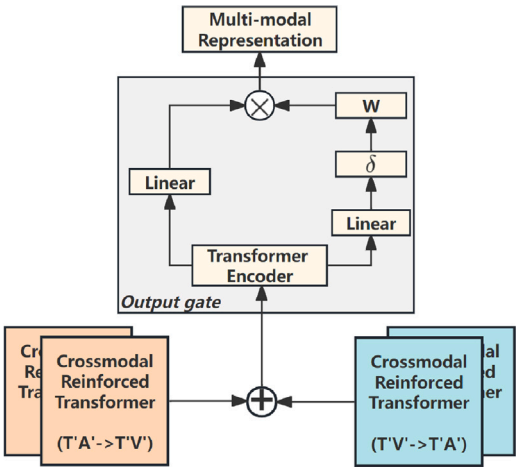


Fig. 5. Multimodal representation output gate.

Assuming that there is a pair of vectors X, denoted as $A = [a_1, a_2, a_3, \ldots .., a_n]$, and $B = [b_1, b2, b_3, \ldots .., b_n]$, the Euclidean distance between them is calculated as:

$$Distance_{euclidean}(A, B) = E(X) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{11}$$

Therefore, if the positive sample set and negative sample set are denoted as $P = [p_1, p_2, p_3, \ldots, p_{2n}]$ and $N = [s_1, s_2, s_3, \ldots, s_{2n^2-2n}]$, then the TCCL loss in the model is calculated as:

$$Distance_{positive} = E(P) = \sum_{j=1}^{2n} E(p_j) \tag{12}$$

$$Distance_{negative} = E(N) = \sum_{k=1}^{2n^2-2n} E(s_j) \tag{13}$$

$$Loss_{TCCL} = -\log \frac{\sum_{k=1}^{2n^2-2n} E(s_j)}{\sum_{j=1}^{2n} E(p_j) + \sum_{k=1}^{2n^2-2n} E(s_j)} \tag{14}$$

3.6. Multimodal representation output gate

In the High-level Fusion process, we concatenate the enhanced text T′ with the video V′ to obtain the bi-modal representation $T'V' \in \mathbb{R}^{L\times(2d_t+2)}$. Similarly, we concatenate the enhanced text T′ with the audio A′ to obtain the bi-modal representation $T'A' \in \mathbb{R}^{L\times(2d_t+2)}$. Both bi-modal representations undergo cross-fusion using CRT (see Eqs. (15) and (16)). In other words, in the T′V′ enhancement process, T′V′ is linearly transformed into K and V, and T′A′ is transformed into Q. In the T′A′ enhancement process, T′A′ is linearly transformed into K and V, and T′V′ is transformed into Q. The two enhanced bi-modal representations are then concatenated to obtain the trimodal representation $M \in \mathbb{R}^{L\times(4d_t+4)}$ (see Eq. (17)).

$$T'A' \rightarrow T'V' = CRT(T'A', T'V') \tag{15}$$

$$T'V' \rightarrow T'A' = CRT(T'V', T'A') \tag{16}$$

$$M = Concat(T'A' \rightarrow T'V'; T'V' \rightarrow T'A') \tag{17}$$

This cross-fusion method captures multimodal sentiment information centred on the text, resulting in a more informative multimodal representation. However, owing to the extensive use of multi-head self-attention in the fusion process, the formed multimodal representation inevitably contains redundant information. Inspired by LSTM gating structures, we design a multimodal feature output gate to address these issues, as illustrated in Fig. 5. Through this gate structure, the model autonomously decides which information is crucial for the output, thereby reducing redundancy. This process involves putting the multimodal fusion representation M into the Transformer encoder to capture the long-term dependencies. The linear layer and sigmoid activation function are then applied to limit the mapped vectors' values to [0,1] as the weights for the multimodal representation output. When weight of the value is closer to 1, the information is more critical. The final output multimodal representation O is obtained as follows:

$$Weight_{output} = Sigmoid(Linear(Encoder(M))) \tag{18}$$

$$O = Weight_{output} \times (Linear(Encoder(M))) \tag{19}$$

In addition, sentiment prediction (see Eq. (20)) is performed using multiple linear layers.

$$pred = Linear(O) \tag{20}$$

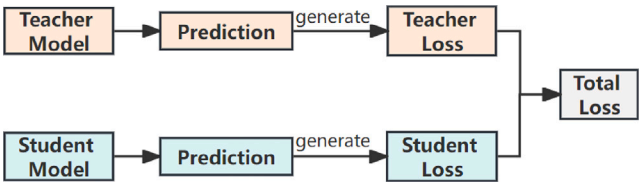


Fig. 6. Knowledge distillation.

3.7. Knowledge distillation in TCHFN

To fully leverage the limited labelled data and enhance model performance, we introduce knowledge distillation, as illustrated in Fig. 6. Specifically, because we set the model parameters after the initial training as the teacher model to guide the subsequent training of the student model, knowledge distillation is not applied during the first training process. Simultaneously, during the subsequent training guided by the teacher model, once the performance of the student model surpasses that of the previous teacher model, the parameters of the student model are saved and updated as a new teacher model to guide the subsequent training of the student model. Consequently, after each training session of the student model, the teacher model is loaded, and the loss value is calculated.

The loss of the student model (see Eq. (21)) equals the task prediction loss (see Eq. (22)) add the TCCL loss value (see Eq. (14)). The calculation method for the teacher model loss is the same as that for the student model's loss.

$$Loss_{student} = (1 - \lambda) \times Loss_{prediction} + \lambda \times Loss_{TCCL} \tag{21}$$

The calculation of the task prediction loss value is as follows:

$$Loss_{prediction} = \frac{1}{N}\sum_{k=1}^{N}(\left|pred^k - y^K\right|) \tag{22}$$

The weighted sum of the teacher and the student model losses (see Eq. (21)) is used as the total loss (see Eq. (23)).

$$Loss_{total} = \alpha \times (Loss_{student}) + (1 - \alpha) \times (Loss_{teacher}) \tag{23}$$

where y represents the true sentiment label, $pred$ is the predicted sentiment label, N is the number of training samples, and α is the parameter of knowledge distillation.

4. Experimental settings

In this section, we first introduce the experimental environment and parameters, followed by the datasets used, evaluation metrics for the experiments, and baseline models.

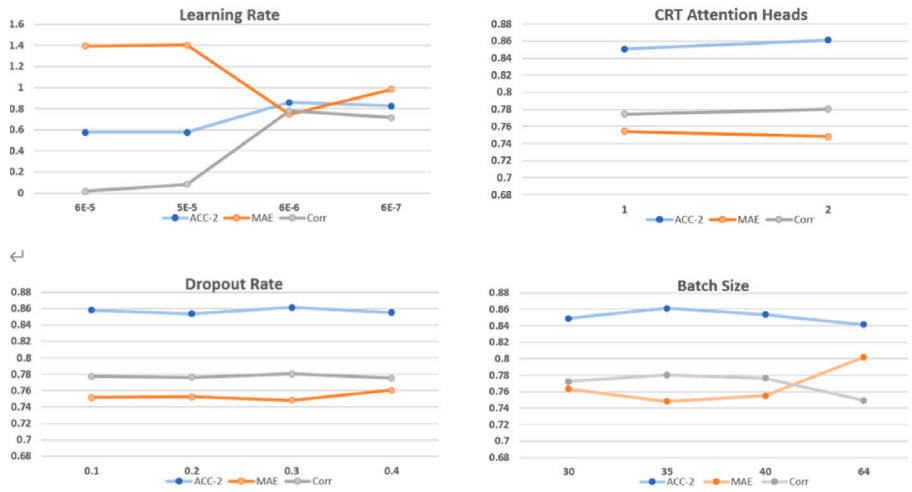


Fig. 7. Results on CMU-MOSI with different learning rates, attention heads, dropout rates, and batch sizes. When changing a single variable, other parameters are fixed as the benchmark set. Mean absolute error (MAE) is lower, the better; ACC-2, Corr is higher, the better. The best performance is achieved when the learning rate, attention heads, dropout rate, and batch size are set to 6e−6, 2, 0.3, and 35, respectively.

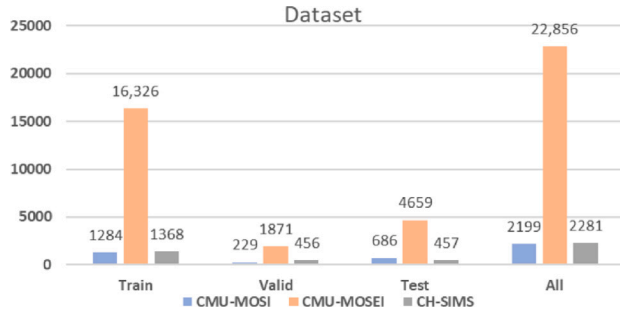


Fig. 8. Dataset statistics in CMU-MOSI, CMU-MOSEI, and CH-SIMS.

4.1. Experimental environment and parameters

We use PyCharm, PyTorch 1.10.1, and Python 3.9. The hardware configuration include an Intel (R) Core (TM) i7-12700F CPU @ 2.10 GHz, a 1 TB hard disc, and a GeForce RTX 3090 GPU, all running on Windows 10. Additionally, we use the Adam optimiser with a learning rate of 6e−6, and the length of the multimodal features is set to 50. CRT cross-modal attention has two heads with two cross-modal blocks. The utterance-level LSTM has one hidden dimension of 1, while the three-modal hidden dimension is 768. The dropout rate is 0.3, and the value is 0.1. For the CMU-MOSI dataset, the batch size is configured as 35; for the CMU-MOSEI dataset, the batch size is specified as 32, and for the CH-SIMS dataset, the batch size is set to 30. Fig. 7 illustrates the impact of varying the values of different parameters on the performance of the model in the CMU-MOSI.

4.2. Datasets

We used three datasets in our experiments: CMU-MOSI, CMU-MOSEI, and CH-SIMS, which are among the most widely used datasets in the current field of MSA. The compositions are shown in Fig. 8.

CMU-MOSI [53]: Curated by Carnegie Mellon University, this dataset consists of 93 videos from 89 speakers on YouTube. The videos were segmented into 2198 emotional video clips. Sentiment intensity scores range from −3 to 3, where −3 represents a strong negative sentiment, 0 is neutral, and 3 represents a strong positive sentiment.

CMU-MOSEI [54]: This dataset is the largest in the field of MSA and contains 22,856 annotated video segments and 250 topics from 1000 speakers. Compared with CMU-MOSI, CMU-MOSEI offers richer video content.

CH-SIMS [55]: Chinese MSA dataset. Using 2281 carefully curated video clips, each sample in the dataset was assigned a sentiment score ranging from −1 to 1. Here, −1 indicates a strong negative sentiment, and 1 represents a strong positive sentiment.

4.3. Evaluation indices

In this study, the classification accuracy is divided into two categories: seven-class classification accuracy (ACC-7) and binary classification accuracy (ACC-2).

ACC-7 ↑: Evaluates the model's accuracy in classifying sentiment into seven categories [−3, 3]. Higher values indicate better performance.

ACC-2 ↑: Measures binary classification accuracy for non-negative/negative (NN/N) and positive/negative (P/N) categories. Higher values indicate better performance.

Additional evaluation metrics include:

F1-Score (**F1** ↑): Considers both precision and recall, subdivided into (NN/N) and (P/N) categories. Higher values indicate better performance.

Mean Absolute Error (**MAE** ↓): Measures the average error between predicted and true sentiment intensities or continuous values. Lower values indicate better performance.

Pearson Correlation (**Corr** ↑): Measures the correlation between predicted sentiment intensity and true sentiment. Higher values indicate a better correlation.

4.4. Baseline

To validate the effectiveness of the TCHFN, we compared it with state-of-the-art models developed in recent years in the field of MSA.

Unimodal: Predicts outputs directly after feature extraction from each modality.

TFN [13]: Learns relationships within and between modalities through triple-Cartesian products.

RAVEN [43]: Enhances multimodal fusion with a bias in non-text information added to the text modality.

MULT [8]: Improved transformer with cross-modal capabilities for multimodal fusion.

GATE [58]: Captures long-term context through self-attention and selectively utilises cross-modal features with a gating mechanism.

Table 1
Results on CMU-MOSI and CMU-MOSEI. We tested two sets of non-negative/negative (left) and positive/negative (right) evaluations for the ACC-2 and F1. The best results are marked in bold.

Model	CMU-MOSEI					CMU-MOSI				
	MAE	Corr	ACC-7	ACC-2	F1	MAE	Corr	ACC-7	ACC-2	F1
V	0.846	0.168	41.59	71.02/62.88	59.07/77.13	1.428	0.071	15.45	40.38/42.23	23.23/59.38
A	0.859	0.090	41.36	71.02/62.85	58.99/77.19	1.413	0.134	15.45	59.47/57.62	44.47/73.00
T	0.664	0.645	45.61	79.05/80.73	79.42/80.88	0.885	0.746	37.46	82.21/82.62	82.16/82.62
TFN[b]	0.573	0.714	51.60	78.50/81.89	78.96/81.74	0.901	0.698	34.90	–/80.20	–/80.70
RAVEN[b]	0.614	0.662	50.00	79.10/–	79.50/–	0.915	0.691	33.20	78.00/–	76.60/–
MULT[a]	0.580	0.703	51.80	–/82.50	–/82.30	0.871	0.698	40.00	–/83.00	–/82.80
GATE[a]	–	–	–	–/81.14	–/78.53	–	–	–	–/83.91	–/81.17
ICCN[a]	0.565	0.713	51.58	–/84.18	–/84.15	0.862	0.714	39.01	–/83.07	–/83.02
MISA[a]	0.555	0.756	52.20	83.60/85.50	83.80/85.30	0.783	0.761	42.30	81.80/83.40	81.70/83.60
MAG-BERT[c]	0.567	0.742	–	81.30/84.80	81.70/84.70	0.778	0.766	–	82.50/84.00	82.40/84.00
MPT[a]	–	–	–	–/82.60	–/82.80	–	–	–	–/ 82.80	–/82.90
TCSP[a]	0.576	0.715	–	–/82.80	–/82.70	0.908	0.710	–	–/80.90	–/81.00
BIMHA[a]	0.562	0.729	52.69	83.19/83.93	83.21/83.64	0.929	0.663	35.86	78.57/80.18	78.55/80.23
MMLATCH[a]	0.582	0.704	52.10	–/82.80	–/82.90	–	–	–	–	–
SUGRM[a]	0.541	0.758	–	83.90/85.00	83.90/85.10	**0.723**	**0.798**	–	82.80/84.50	82.80/84.50
MUTA[a]	0.544	0.760	–	82.40/85.00	82.70/84.90	0.730	0.793	–	83.10/85.00	83.00/85.00
TCHFN	**0.538**	**0.770**	**53.19**	**84.01/86.27**	**84.14/86.48**	0.748	0.780	**44.75**	**85.57/86.13**	**85.41/86.31**

[a] Indicates results from the original papers.
[b] Indicates results from [56].
[c] Indicates results from [57].

Table 2
Results on CH-SIMS.

Model	CH-SIMS			
	MAE	Corr	ACC-2	F1
TFN[b]	0.488	0.496	75.27	75.56
MulT[b]	0.485	0.504	75.62	75.84
MISA[b]	0.472	0.542	75.49	75.85
MAG-BERT[b]	0.553	0.242	71.43	63.68
ConKI[a]	0.454	0.542	77.94	78.17
MUTA[a]	0.428	0.594	79.17	79.28
VLP2MSA[a]	0.413	0.604	79.43	79.26
TCHFN	**0.329**	**0.734**	**80.27**	**80.44**

[a] Indicates results from the original papers.
[b] Indicates results from [45].

ICCN [14]: Analyses hidden relationships between modalities using deep canonical correlation analysis.

MISA [10]: Maps multimodal spaces into separate spaces to identify and effectively fuse shared and private features.

MAG-BERT [7]: Integrates non-text features with text features using MAG in BERT training.

MPT [59]: Compresses long sequences into shorter hidden state sequences using a phased transformer.

TCSP [47]: Designs two distinct cross-modal transformers to obtain shared and private features.

BIMHA [56]: Fuses acoustic-visual, acoustic-textual, and visual-textual features through a fusion network based on multi-head attention.

MMLATCH [60]: Extracts high-level representations for each modality to block sensory inputs for top-down cross-modal interactions.

SUGRM [57]: Recalibrates each unimodal representation to optimise the learning representation of multimodal.

ConKI [45]: Addresses the impact of domain-specific knowledge on MSA through Knowledge Injection.

MUTA [61]: Enhances multimodal fusion representations through increased inter-class distance and reduced intra-class feature distance in the fusion process.

VLP2MSA [62]: Enhances the importance of visual modality in MSA by introducing of an inter-frame hybrid transformer.

5. Experimental results and analysis

5.1. Quantitative results

The comparative results of MSA on CMU-MOSI and CMU-MOSEI are presented in Table 1. The current research provides datasets for both aligned and non-aligned versions at the word level [40]. Generally, the aligned datasets perform better, but we opt for a non-aligned version to better reflect real-world scenarios and present a more challenging task.

On the left-hand side of Table 1, it is evident that our model achieves the best results across all metrics on the CMU-MOSEI dataset. The ACC-2 values for non-negative/negative (NN/N) and positive/negative (P/N) are 84.01% and 86.27%, respectively, far surpassing the metrics of other baseline models. The F1-scores for (NN/N) and (P/N) are 84.14% and 86.48%, respectively, significantly outperforming the other baseline models. The ACC-7 is 53.19%, which is 0.5% higher than that of the best-performing baseline model BIMHA. The Corr value is 0.770, surpassing that of the best-performing baseline model MUTA by 0.01. Meanwhile, the MAE value is 0.538, which is 0.003 lower than that of the best-performing baseline model SUGRM.

Moving to the right-side of Table 1, it is evident that our model achieves the best results for the ACC-7, ACC-2, and F1-scores on the CMU-MOSI dataset. The ACC-2 (NN/N) and F1-score (NN/N) exceed the best baseline model MUTA by 2.47% and 2.41%, respectively. The ACC-7 is 2.45% higher than that of the best baseline MISA model. The MAE and Corr values outperform most baseline models but are slightly weaker than those of SUGRM and MUTA. However, our model achieves the best MAE and Corr values on the CMU-MOSEI dataset, possibly because the CMU-MOSI dataset size is insufficient to train our model effectively for the regression tasks. Additionally, our model's ACC-7, ACC-2, and F1-score are superior to those of SUGRM and MUTA, suggesting that, overall, our model performs better on the CMU-MOSI dataset than SUGRM and MUTA.

Furthermore, by examining Table 2, it is evident that our model achieves optimal performance across all metrics on the CH-SIMS dataset. Specifically, the metrics ACC-2, MAE, and Corr metrics outperform the best-performing baseline model VLP2MSA by 0.84%, 0.084, and 0.13, respectively. Additionally, the F1-score surpasses that of the baseline MUTA model by 1.16%.

In conclusion, the comparative experiments indicate that our model demonstrates the best performance across the three multimodal datasets. MAE and Corr are commonly used indicators in regression

Table 3
Ablation experiment on CMU-MOSI dataset. '–' denotes removing some submodules from the whole TCHFN.

Description	Model	MAE	Corr	ACC-7	ACC-2	F1
Effect of text modality	-V	0.829	0.746	36.73	84.54/85.06	84.35/85.28
	-A	0.849	0.727	36.44	83.38/83.68	83.08/84.01
	-T	1.425	0.127	14.86	52.04/53.05	48.57/56.73
	VCHFN	0.798	0.766	41.83	84.69/85.06	84.47/85.29
	ACHFN	0.757	0.774	42.71	84.69/85.36	84.55/85.53
Effect of text enhancement module	-enhance_U	0.787	0.768	43.15	85.28/85.98	85.11/86.17
	-enhance_T*	0.772	0.765	44.16	85.13/85.67	84.92/85.89
	-enhance_text	0.794	0.762	39.21	84.55/85.37	84.49/85.46
Effect of non-text enhancement module	-TCCL	0.749	0.778	44.02	85.27/85.82	85.12/86.00
	-enhance_nontext	0.818	0.754	40.81	84.83/85.06	84.52/85.39
Effect of multimodal output gate module	-output_gate	0.784	0.779	42.13	85.27/85.82	85.05/86.06
Overall modal	**TCHFN**	**0.748**	**0.780**	**44.75**	**85.57/86.13**	**85.41/86.31**

Table 4
Compared predictions of TCHFN, ACHFN, and VCHFN. By preserving the parameters of the TCHFN, ACHFN, and VCHFN models and passing the following four sets of data to each model, we obtain the corresponding sentiment prediction values.

Text	Visual	Audio	Truth	TCHFN	ACHFN	VCHFN
Because I truly love an action flick action comedy flick even better right	Laugh	Cheerful tone	**1.6**	**1.6978**	1.3484	1.3848
So um I thought she did greatwith this movie	Smile	Slightly rising tone	**2.0**	**1.9822**	1.4283	1.4786
The first problem is that its all told in flashback from a point just before Katrina hits new Orleans	Glance	Peaceful	**−1.8**	**−1.8197**	−1.6765	−1.6160
It just wasnt it di did not work for me	Frown	Narrative	**−2.0**	**−1.9901**	−1.8222	−1.7693

tasks, and the optimal performances of CMU-MOSEI and CH-SIMS indicate that our model can accurately capture linear relationships with real emotional values. The ACC-7, ACC-2, and F1-score to evaluate the model. Despite its challenges, our model achieves optimal results across three multimodal datasets, illustrating the effectiveness of our model of classification tasks. These results is achieved by primarily focusing on textual modality, utilising gate structures to minimise information redundancy within multimodal representations, and incorporating knowledge distillation to fully exploit the limited multimodal label data. Additionally, the analysis of the three unimodalities on the CMU-MOSI and CMU-MOSEI datasets shows that text modality outperforms audio and video in all metrics, emphasising the crucial role of text modality in unimodal sentiment analysis. A comparison of our model's predictions with those of each unimodality reveals that the multimodal fusion representations outperform the unimodal representations, indicating the model's ability to combine information from all three modalities for sentiment prediction.

5.2. Ablation studies

Effect of text modality as the key modality: In the unimodal predictions presented in rows 1–3 of Table 1, we validate the role of text modality in unimodal sentiment analysis. To investigate its impact on MSA further, we conduct ablation experiments specifically focusing on the text modality. This involves evaluating the performance when the model lacks video, audio, and text individually, as well as the effect of fusion with video and audio as the central modalities (VCHFN and ACHFN). As observed in rows 1–3 of Table 3, removing the text modality has a more pronounced impact than eliminating the visual or acoustic modalities. Specifically, the F1-scores for (NN/N) and (P/N) decrease by 36.84% and 29.58%, respectively. Performance also significantly drops when using video or audio as the central modality compared to text. Additionally, as evident from Table 4, when provide with the same input data, the predicted label values by the TCHFN are closer to the true label values than those predicted by the VCHFN and ACHFN. Combining this analysis with the information from rows 4–5 of Table 3, it can be concluded that the predictive performance of the TCHFN is significantly superior to that of the ACHFN and VCHFN. This underscores the crucial importance of the text modality in MSA. Therefore, the text modality emerges as the most crucial modality in MSA.

Effect of text enhancement module: To validate the effectiveness of the text enhancement module in our TCHFN model, we conduct Several comparative experiments. Specifically, we examine the impact of removing various components of the text enhancement module, including utterance-level enhancement, cross-modal enhancement, and the entire text enhancement module, where the initial text vector is directly integrated after BERT processing. When the model loses utterance-level enhancement but retains cross-model enhancement, there is a noticeable, although not drastic, decrease in performance. The data in rows 6–8 of Table 3 indicate that ACC2 (NN/N) and F1 (NN/N) decrease by 0.29% and 0.3%, respectively. This indicates that utterance-level enhancement contributes to performance, but it is not as critical as cross-model enhancement. In contrast, when the model loses cross-modal enhancement while maintaining utterance-level enhancement, the performance decrease is more substantial. ACC2 (NN/N) and F1 (NN/N) decrease by 0.44% and 0.49%, respectively. This suggests that cross-modal enhancement, which integrates information from visual and acoustic modalities, plays a more vital role in improving model performance compared to utterance-level enhancement. Removing the entire text enhancement module and using text presentation processed through BERT for later fusion results in the most significant performance drop. ACC2 (NN/N) and F1 (NN/N) decrease by 1.02% and 0.92%, respectively. This underscores the importance of the text enhancement module, particularly the integration of cross-modal information, for achieving superior model performance. This suggests that information from visual and acoustic modalities significantly complements text modality information, and the inclusion of long-distance dependency information is indispensable for improved fusion.

Effect of non-text enhancement module: To assess the impact of the non-text enhancement module, we first remove the TCCL module from the model. Since contrastive learning facilitates the alignment of non-text embeddings and helps distinguish relationships between samples, removing the TCCL leads to a performance decline. As presented in rows 9–10 of Table 3, ACC2 (NN/N) and F1 (NN/N) decrease by 0.3% and 0.29%, respectively. Moreover, as shown in Fig. 9, before applying TCCL, the data distribution of the three modalities is highly scattered. With the use of TCCL, non-text modal data tends to cluster

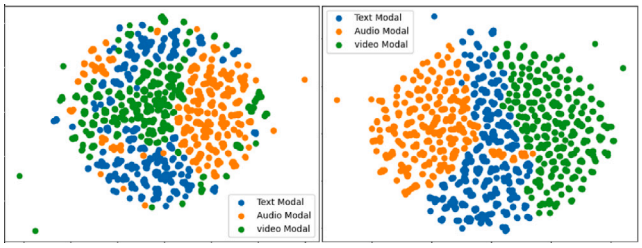


Fig. 9. Comparison of TCCL effects chart on CMU-MOSI dataset using T-SNE projections. The left chart illustrates the distribution of the trimodal data when the model does not utilise TCCL, while the right chart illustrates the distribution after implementing TCCL. Blue dots represent textual data, orange dots represent acoustic data, and green dots represent visual data.

Table 5
Knowledge distillation parameter comparison experiment on CMU-MOSI dataset.

Description	MAE	Corr	ACC-7	ACC-2	F1
α=0.2	0.823	0.758	37.17	85.28/85.37	84.99/85.65
α=0.3	0.807	0.756	41.40	85.56/85.97	85.36/86.19
α=0.4	0.750	0.779	44.17	84.99/85.82	84.87/85.98
α=0.5	0.758	0.776	43.73	84.99/85.51	84.81/85.71
α=0.6	0.756	0.779	44.02	84.98/85.57	84.91/85.77
α=**0.7**	**0.748**	**0.780**	**44.75**	**85.57/86.13**	**85.41**/86.31
α=0.8	0.791	0.769	42.41	85.42/86.12	85.22/**86.35**
α=1.0	0.803	0.766	41.40	85.13/85.82	84.94/86.03

around the text modality data centres. This observation underscores the efficacy of TCCL in aligning non-text modalities with text. Subsequently, we remove the entire non-text enhancement module. This resulted in the most noticeable decrease in performance. The ACC2 (NN/N) and F1 (NN/N) decrease by 0.74% and 0.89%, respectively. Therefore, the non-text enhancement module contributes to improving model performance by aligning and learning from non-text modalities using text representations.

Effect of multimodal output gate module: As presented in row 11 of Table 3, upon removing the multimodal output gate, all performance metrics exhibit a decline, indicating that the multimodal output gate module effectively eliminates redundant information during the fusion process, thereby enhancing the overall model performance.

Effect of knowledge distillation module: As presented in Table 5, we evaluate the effectiveness of the knowledge distillation module by adjusting its hyperparameter α. When the α is set to 1.0, indicating no usage of the teacher model, it is equivalent to omitting the knowledge distillation module. As the α decreases, the model increasingly relies on the teacher model, and the loss function approaches that of the teacher model. However, excessive dependence on the prior knowledge of the teacher model can make it challenging for the student model to surpass the teacher model. Therefore, selecting an appropriate α is crucial to fully leveraging the teacher model's role. Through experiments, we observe that the performance metrics are optimal when α is set to 0.7, except for a slightly lower F1-score (P/N) compared to when α is 0.8. Furthermore, performance improved compare to not using the knowledge distillation module (α = 1.0), indicating that knowledge distillation can assist the model in learning more knowledge from the limited dataset, thereby enhancing overall performance.

6. Conclusion and prospect

This study introduces an innovative TCHFN designed for MSA tasks. The model, centred on text, comprises two crucial steps: low- fusion and high-level fusion. In the low-level fusion stage, combining the CRT with text-sentence-level representations facilitated the text enhancement network. Simultaneously, the integration of CRT with TCCL contributes to a non-text-enhancement network. This approach effectively consolidates information from the text, video, and audio modalities. In the high-level fusion stage, the CRT further establishes interactions between text, video, and audio multimodal representations. Using a multimodal representation output gate to reduce redundancy, the model obtains more effective multimodal fusion representations for sentiment prediction. Furthermore, to improve model performance, knowledge distillation is introduced, which enabled the model to learn additional emotional information during the training process. Compared with previous works, TCHFN achieves competitive results in MSA tasks on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, demonstrating its capability to learn superior multimodal representations. Through ablation studies, the importance of text modality is highlighted, emphasising the effectiveness of the text-enhancement network with CRT and utterance-level representations, non-text enhancement network with CRT and TCCL, and multimodal output gate module. The effectiveness of the knowledge distillation module is validated by varying the hyperparameter values. In future work, we will focus on refining the core role of the text modality and exploring methods for identifying similarities and dissimilarities between text and non-text modalities. The aim is to simplify the model while obtaining richer multimodal fusion representations to enhance task accuracy.

CRediT authorship contribution statement

Jingming Hou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nazlia Omar:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation. **Sabrina Tiun:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision. **Saidah Saad:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Qian He:** Writing – review & editing, Writing – original draft, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the sources of the datasets in my paper.

Acknowledgments

This work is supported in part by Universiti Kebangsaan Malaysia under Grant TAP-K007009, in part by the National Natural Science Foundation of China under Grant 62162018 and Grant 61861013, in part by Guangxi Natural Science Foundation, China 2023JJD170008.

References

[1] R. Grewal, S. Gupta, R. Hamilton, Marketing insights from multimedia data: text, image, audio, and video, J. Mar. Res. 58 (6) (2021) 1025–1033.
[2] A. Pandey, D.K. Vishwakarma, Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey, Appl. Soft Comput. (2023) 111206.
[3] S. Lai, X. Hu, H. Xu, Z. Ren, Z. Liu, Multimodal sentiment analysis: A survey, Displays (2023) 102563.
[4] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: 2017 IEEE International Conference on Data Mining, ICDM, IEEE, 2017, pp. 1033–1038.
[5] N. Xu, W. Mao, Multisentinet: A deep semantic network for multimodal sentiment analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 2399–2402.

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

[7] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2020, NIH Public Access, 2020, p. 2359.

[8] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.

[9] J. He, S. Mai, H. Hu, A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis, IEEE Signal Process. Lett. 28 (2021) 992–996.

[10] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.

[11] J. Yang, Y. Yu, D. Niu, W. Guo, Y. Xu, ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 7617–7630.

[12] Z. Li, Y. Zhou, W. Zhang, Y. Liu, C. Yang, Z. Lian, S. Hu, AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 7136–7146.

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

[14] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8992–8999.

[15] B. Min, H. Ross, E. Sulem, A.P.B. Veyseh, T.H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2) (2023) 1–40.

[16] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, X. Huang, TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis, Knowl.-Based Syst. 269 (2023) 110502.

[17] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, X. Luo, TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis, Pattern Recognit. 136 (2023) 109259.

[18] S. Mai, Y. Zeng, H. Hu, Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations, IEEE Trans. Multimed. (2022).

[19] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).

[20] N.A. Osman, S.A. Mohd Noah, M. Darwich, M. Mohd, Integrating contextual sentiment analysis in collaborative recommender systems, PLoS One 16 (3) (2021) e0248695.

[21] S.K. Hamed, M.J. Ab Aziz, M.R. Yaakub, Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments, Sensors 23 (4) (2023) 1748.

[22] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: Approaches, challenges and trends, Knowl.-Based Syst. 226 (2021) 107134.

[23] E. Sukawai, N. Omar, Corpus development for Malay sentiment analysis using semi supervised approach, Asia-Pac. J. Inf. Technol. Multimedia 9 (01) (2020) 94–109.

[24] M.M. Kabir, Z.A. Othman, M.R. Yaakub, S. Tiun, Hybrid syntax dependency with lexicon and logistic regression for aspect-based sentiment analysis, Int. J. Adv. Comput. Sci. Appl. 14 (10) (2023).

[25] H. Li, Q. Chen, Z. Zhong, R. Gong, G. Han, E-word of mouth sentiment analysis for user behavior studies, Inf. Process. Manage. 59 (1) (2022) 102784.

[26] M. Darwich, S.A.M. Noah, N. Omar, Deriving the sentiment polarity of term senses using dual-step context-aware in-gloss matching, Inf. Process. Manage. 57 (6) (2020) 102273.

[27] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: Applications, challenges and trends, Neurocomputing 408 (2020) 189–215.

[28] Priyanka, D. Kumar, Decision tree classifier: a detailed survey, Int. J. Inf. Decis. Sci. 12 (3) (2020) 246–269.

[29] C. Sutton, A. McCallum, et al., An introduction to conditional random fields, Found. Trends® Mach. Learn. 4 (4) (2012) 267–373.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[32] S. Mai, S. Xing, H. Hu, Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 1424–1437.

[33] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Syst. Appl. 173 (2021) 114683.

[34] R.A. Patamia, W. Jin, K.N. Acheampong, K. Sarpong, E.K. Tenagyei, Transformer based multimodal speech emotion recognition with improved neural networks, in: 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning, PRML, IEEE, 2021, pp. 195–203.

[35] X. Song, K. Huang, W. Gao, Facelistener: Recognizing human facial expressions via acoustic sensing on commodity headphones, in: 2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN, IEEE, 2022, pp. 145–157.

[36] L. Wu, M. Qi, M. Jian, H. Zhang, Visual sentiment analysis by combining global and local information, Neural Process. Lett. 51 (2020) 2063–2075.

[37] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, Q. Sun, Feature pyramid transformer, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, 2020, pp. 323–339.

[38] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325.

[39] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Sign. Proces. 14 (3) (2020) 478–493.

[40] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10790–10797.

[41] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[42] J. Zeng, J. Zhou, T. Liu, Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 2924–2934.

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

[44] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, Y. Li, Unimse: Towards unified multimodal sentiment analysis and emotion recognition, 2022, arXiv preprint arXiv:2211.11256.

[45] Y. Yu, M. Zhao, S.-a. Qi, F. Sun, B. Wang, W. Guo, X. Wang, L. Yang, D. Niu, ConKI: Contrastive knowledge injection for multimodal sentiment analysis, 2023, arXiv preprint arXiv:2306.15796.

[46] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, W. Kong, Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5301–5311.

[47] Y. Wu, Z. Lin, Y. Zhao, B. Qin, L.-N. Zhu, A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4730–4738.

[48] K. Kim, S. Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.

[49] F. Lv, X. Chen, Y. Huang, L. Duan, G. Lin, Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2554–2562.

[50] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[51] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3967–3976.

[52] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2604–2613.

[53] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, arXiv preprint arXiv:1606.06259.

[54] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, 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), 2018, pp. 2236–2246.

[55] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, 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, 2020, pp. 3718–3727.

[56] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, Y. Huang, Video sentiment analysis with bimodal information-augmented multi-head attention, Knowl.-Based Syst. 235 (2022) 107676.

[57] Y. Hwang, J.-H. Kim, Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 35–46.

[58] A. Kumar, J. Vepa, Gated mechanism for attention based multi modal sentiment analysis, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 4477–4481.

[59] J. Cheng, I. Fostiropoulos, B. Boehm, M. Soleymani, Multimodal phased transformer for sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 2447–2458.

[60] G. Paraskevopoulos, E. Georgiou, A. Potamianos, Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 4573–4577.

[61] Z. Tang, Q. Xiao, X. Zhou, Y. Li, C. Chen, K. Li, Learning discriminative multi-relation representations for multimodal sentiment analysis, Inform. Sci. 641 (2023) 119125.

[62] G. Yi, C. Fan, K. Zhu, Z. Lv, S. Liang, Z. Wen, G. Pei, T. Li, J. Tao, VLP2MSA: Expanding vision-language pre-training to multimodal sentiment analysis, Knowl.-Based Syst. 283 (2024) 111136.