Information Fusion 100 (2023) 101958

Contents lists available at ScienceDirect

Information Fusion

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

SKEAFN: Sentiment Knowledge Enhanced Attention Fusion Network for multimodal sentiment analysis

Chuanbo Zhu [a], Min Chen [b,c], Sheng Zhang [a], Chao Sun [a], Han Liang [a], Yifan Liu [a], Jincai Chen [a,b,d,*]

[a] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China
[b] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China
[c] Embedded and Pervasive Computing (EPIC) Lab, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China
[d] Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China, Wuhan, 430074, Hubei, China

ARTICLE INFO

Keywords:
Multi-view learning
Multiple feature fusion
Multimodal sentiment analysis
External knowledge
Multi-head attention

ABSTRACT

Multimodal sentiment analysis is an active research field that aims to recognize the user's sentiment information from multimodal data. The primary challenge in this field is to develop a high-quality fusion framework that effectively addresses the heterogeneity among different modalities. However, prior research has primarily concentrated on intermodal interactions while neglecting the semantic sentiment information conveyed by words in the text modality. In this paper, we propose the Sentiment Knowledge Enhanced Attention Fusion Network (SKEAFN), a novel end-to-end fusion network that enhances multimodal fusion by incorporating additional sentiment knowledge representations from an external knowledge base. Firstly, we construct an external knowledge enhancement module to acquire additional representations for the text modality. Then, we design a text-guided interaction module that facilitates the interaction between text and the visual/acoustic modality. Finally, we propose a feature-wised attention fusion module that achieves multimodal fusion by dynamically adjusting the weights of the additional and each modality's representations. We evaluate our method on three challenging multimodal sentiment analysis datasets: CMU-MOSI, CMU-MOSEI, and Twitter2019. The experiment results demonstrate that our model significantly outperforms the state-of-the-art models. The source code is publicly available at https://github.com/doubibobo/SKEAFN.

1. Introduction

With the rapid advancement of social media technology, there has been a significant surge in the volume of sentiment information available on the Internet. Analyzing and processing this vast amount of data poses considerable challenges [1]. Multimodal Sentiment Analysis (MSA), a crucial technology in intelligent human–computer interaction, has garnered substantial attention [2]. In contrast to unimodal Sentiment Analysis (SA), MSA focuses on extracting the sentiment tendencies of users from multimodal signals, encompassing text, visual, acoustic, and other modalities [3]. Its applications span various domains, including online shopping, medical services, depression detection, fake news identification, and other fields [4].

Generally, different modalities can provide complementary sentiment semantic information to one another [5]. However, multimodal signals used in MSA are often heterogeneous [6]. For instance, the text modality consists of multiple discrete words with specific meanings, while the visual/acoustic modality consists of continuous digital signals. Consequently, developing a high-quality fusion framework to integrate the heterogeneous sentiment information from multiple modalities effectively has become a key challenge in MSA. To address this challenge, researchers have proposed a large number of approaches. Some approaches aim to incorporate supplementary information into multimodal data to achieve similar performance gains from SA to MSA. Examples of this additional information encompass fine-grained objects and their corresponding regions in the visual modality [7,8], shared information on the emotional dimension [9,10], and physical indicators like ECG and RSP signals [11], among others. Alternatively, other approaches concentrate on constructing diverse fusion architectures to facilitate interactions within and between modalities. These fusion approaches can be broadly classified as tensor-based [12–14], graph-based [15–17], and seq2seq-based [3,18–20] models. It is worth mentioning that these fusion models often incorporate deep learning

* Corresponding author at: Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China.
E-mail addresses: chuanbo_zhu@hust.edu.cn (C. Zhu), minchen2020@hust.edu.cn (M. Chen), zhangmonkey@hust.edu.cn (S. Zhang), chaosun@hust.edu.cn (C. Sun), hanliang@hust.edu.cn (H. Liang), liu_yi_fan@hust.edu.cn (Y. Liu), jcchen@hust.edu.cn (J. Chen).

https://doi.org/10.1016/j.inffus.2023.101958
Received 28 February 2023; Received in revised form 26 July 2023; Accepted 28 July 2023
Available online 2 August 2023

techniques such as transfer learning [21], contrastive learning [5], semi-supervised learning [21,22], and multi-task learning [9].

However, previous studies have overlooked the explicit sentiment semantic information conveyed by the words in the text modality. For example, consider the phrase "It was horrible as the smoke and bad service was top notch". The words "horrible" and "bad" clearly express negative sentiment. Conversely, multimodal tweets containing hashtags such as "#bestoftheday" and "#stillthebest" typically exhibit positive sentiment tendencies because of the word "best". Moreover, besides adjectives, nouns and verbs also carry explicit sentiment tendencies, such as "laugh", "love", "hate", and so on. This explicit sentiment information in words can guide the multimodal fusion process of the MSA model, leading to improved performance. Furthermore, there are occasional implicit connections between modalities. For instance, tweets employed for text-image sarcasm detection often feature text-rich images. This external textual information in the visual modality can be extracted and used as an additional modality for the MSA task. However, this aspect has been disregarded in previous works, thereby constraining the model's performance.

In this paper, we propose the Sentiment Knowledge Enhanced Attention Fusion Network (SKEAFN), which incorporates the sentiment semantic information of words to enhance the multimodal fusion process. The SKEAFN comprises three core units: the text-guided interaction (TGI) module, the external knowledge enhancement (EKE) module, and the feature-wised attention fusion (FWAF) module. In the TGI module, the representation dimension for each modality is unified through the previous modality encoder module, enabling the direct application of the self-attention mechanism [19,23,24] to crossmodal inputs. This crossmodal attention facilitates the understanding of the relationship between text and non-verbal modalities through crossmodal interaction. Specifically, at each layer of the TGI module, we utilize text representation as the query input and non-verbal representations as the inputs for key and value, respectively. In this way, the text modality can guide our model to focus on the relevant regions of non-verbal modalities during the fusion process. Importantly, we incorporate optical character recognition (OCR) results into the multimodal sarcasm detection task, thereby enriching the textual information available.

The EKE module incorporates pre-extracted sentiment knowledge by constructing a sentiment knowledge graph specifically for the text modality. In the graph, each node represents a word associated with specific sensitivity and emotional knowledge derived from SenticNet [25,26], while each edge indicates the relationship between the respective words. Utilizing this graph, we compute graph embeddings to generate graph-level sentiment knowledge embeddings for each utterance. The resulting aggregated embedding, combined with the sentiment semantic information of external words, enhances the joint representation during the multimodal fusion process.

Regarding the FWAF module, we initially integrate the representations of four modalities through a straightforward vector concatenation operation. Subsequently, we devise a feature-wised attention mechanism to dynamically adjust the weights for each dimension in the representation of each modality. This approach allows us to establish correlations between each representation dimension and identify crucial features to enhance the multimodal fusion process. The resulting joint representation from this module is ultimately employed to generate the sentiment prediction.

We evaluate the SKEAFN on two subtasks of MSA: multimodal sentiment intensity prediction and multimodal sarcasm detection. Three public datasets are used: CMU-MOSI [27], CMU-MOSEI [28], and Twitter2019 [7]. The experimental results demonstrate that the SKEAFN outperforms the previous state-of-the-art across nearly all metrics. Furthermore, the ablation study and further analysis prove the effectiveness of our proposed SKEAFN.

Our main contributions to this paper can be summarized as follows:

- We propose the SKEAFN model for MSA by introducing external sentiment knowledge to enhance multimodal fusion.
- To extract external explicit sentiment knowledge from text modality, we design an additional sentiment knowledge generation module by constructing a sentiment knowledge graph and performing graph computing for the MSA task.
- To obtain sentiment information from acoustic and visual modality, we design a text-guided interaction module by generalizing the self-attention mechanism to crossmodal inputs and capturing complementary non-verbal information according to text modality.
- To measure the importance of each modality in the fusion process, we design a feature-wised attention module to set the weights of each modality's representation dynamically.
- Extensive experiments on three public datasets indicate that our model can efficiently use external sentiment knowledge to improve performance and achieve a new state-of-the-art result for the MSA task.

2. Related work

2.1. Multimodal Sentiment Analysis

The MSA task has gained increased interest in recent years. It mainly focuses on mining human sentiments from complex multimedia data composed of text, visual, acoustic, etc. Before fusing information from multiple modalities (*multimodal fusion*), researchers first need to represent heterogeneous multimodal data (*multimodal representation*).

For multimodal representation, the core issue is how to construct representations for heterogeneous modalities. For each modality, approaches have been extensively studied, from hand-designed for specific applications to data-driven [2]. For example, the most popular way to represent a sentence is through data-driven large-scale pre-trained models, such as BERT [29], RoBERTa [30], etc. Similarly, most images are currently represented using descriptions learned from the Convolutional Neural Network (CNN) [31–33]. However, researchers prefer to use feature extraction tools specifically designed for MSA to obtain facial sentiment representations, such as Facet and OpenFace [34]. While in the audio domain, hand-designed acoustic features such as Mel Frequency Cepstral Coefficients (MFCC) are widely used. The multimodal representation builds on unimodal representations by involving simple concatenation or complex Deep Neural Networks (DNN) [35–38]. Some works also commit to enhancing multimodal representation by introducing sentiment information from other latitudes. MultiSentiNet [39] extracts deeper visual semantic features by identifying extra objects and scenes in an image. IIMI-MMSD [40] studies the representation of hashtags with contrasting information to the text. D&R Net [41] utilizes Adjective-noun Pairs (ANPs) to introduce additional contextual information to the input images. Multimodal Sarcasm Target Identification (STI) [8], a subtask of MSA, deserves a further understanding of sarcasm in depth by adding missing sarcasm targets of text and images. The Self-MM [22] generates unimodal labels by the self-supervised method and introduces three unimodal subtasks to auxiliary multimodal representation. The CMKT-SSL [21] leverages unlabeled audio–visual data based on transfer and semi-supervised learning to aid in learning speech representation in Speech Emotion Recognition (SER). The DictABSA [42] incorporates external noun-explanation knowledge from Oxford Dictionary to augment the sentiment polarity identification capability of Aspect-based Sentiment Analysis (ABSA) approaches. The Rce-KGQA [43] infuses the implicit relational chain knowledge stored in structured Knowledge Graph (KG) for multi-hop Knowledge Graph Question Answering (KGQA).

For multimodal fusion, previous approaches can be broadly categorized as follows:

Tensor-based models. Different from simple feature concatenation, such models utilize tensor operations to model correlations among multiple modalities. Tensor Fusion Network (TFN) [14] transforms three modalities features into a 3D-Tensor by performing a tensor

outer product operation. However, the computational complexity for the 3D-Tensor would increase exponentially as the feature dimension increases. To address this issue, Low-Rank Multimodal Fusion (LMF) [13] applies low-rank decomposition to reduce the complexity of the high-order 3D tensor. Dual-LMF [12] further fuses modalities features across the time domain by performing dimension reduction on the temporal dimension. Besides, the Hierarchical Polynomial Fusion Network (HPFN) [44] constructs a polynomial tensor pooling (PTP) block to integrate multimodal features, with recursively transmitting local correlations into global ones.

Graph-based models. Graph Neural Networks (GNN) integrate heterogeneous multimodal data into graphs and learn complex associations within and between modalities, which can be used to acquire better multimodal representations and generate more accurate sentiment prediction results. The Adversarial Representation Graph Fusion model (ARGF) [15] builds the encoded multimodal representation fusion using a hierarchical graph neural network and leverages adversarial training to reduce the modality distribution gap. The Multimodal Graph network [16] transforms the unaligned sequences into a graph and devises graph convolution and pooling algorithms to learn the longer intra- and inter-modal temporal dependency. The Multi-channel Attentive Graph Convolutional Network (MAGCN) [17] exploits densely connected graph convolutional networks to learn inter-modality dynamics and utilizes multi-head self-attention to merge sentimental knowledge into inter-modality feature representations.

Seq2Seq-based models. Inspired by the approaches in natural language processing, sequence-to-sequence learning methods for capturing the modalities interactions over time steps are introduced to the MSA task. The MultiSentiNet model [39] uses Long Short-Term Memory (LSTM) with visual feature-guided attention to fuse multimodal information. The MTL model [9] employs three bi-directional Gated Recurrent Unit (GRU) networks to capture the contextual information for each modality and jointly performs sentiment and emotion analysis by multi-task learning. Although LSTM, GRU, and their variants can memorize long-term dependencies, they cannot dynamically focus on important information in sequences and between modalities. To model the fine-grained interaction of multiple modalities, Seq2Seq-based models with attention mechanisms are proposed [20,45–47]. Recurrent Attended Variation Embedding Network (RAVEN) [48] uses modality-specific attention gates to shift nonverbal cues' word representations dynamically. The Multimodal Adaptation Gate (MAG) module [49] allows the large pre-trained model to accept multimodal nonverbal data during fine-tuning. Cross-Modal BERT (CM-BERT) [19] utilizes masked multimodal attention to adjust the weight of words in the processing of multimodal fusion. With the success of Transformer in the field of machine translation tasks, the self-attention mechanism for sequence-to-sequence learning becomes increasingly popular. The Multimodal Transformer (MulT) [23] latently adapts sequences with the core directional pairwise crossmodal attention from one modality to another. The Bi-Bimodal Fusion Network (BBFN) [18] performs a fusion of relevance increment and separation of difference increment on pairwise modality representations. Moreover, the BIMHA [3] captures interactions among modalities through a multi-head attention-based fusion network.

Inspired by the above works, we introduce the latitude of explicit sentiment semantic information in words to enhance multimodal representation for the MSA task and design a Seq2Seq-based text-guided interaction model with an attention mechanism to fuse multimodal information.

2.2. Sentiment knowledge base

An essential way of obtaining explicit sentiment information in words is to develop concept-level sentiment knowledge bases. Researchers have recently built a series of sentiment dictionaries based on linguistic knowledge and machine-learning techniques. The SentiWordNet [50] is based on the quantitative analysis of the glosses associated with synsets in WordNet [51], and each synset is associated with three scores of objective, positive, and negative. These three scores are obtained by eight ternary classifiers built on vectorial term representations. The scale of the SentiWordNet is 1000 synsets in the original version and expands as the upgrade of WordNet [52]. Unlike SentiWordNet, SenticNet [53,54] focuses more on commonsense than syntax-based synsets sentiment mining. It assumes that relative distances between concepts in AffectiveSpace are directly proportional to their polarity degree difference. By integrating the techniques of commonsense reasoning, emotion classification, and ontologies, the scale of the SenticNet reaches more than 5700 polarity concepts. It also expands with the development of AI, SenticNet 6 [25] integrates logical reasoning within deep learning architectures to extend the size into 200,000 concepts. SenticNet 7 [26] leverages both subsymbolic and symbolic AI to perform a wholly interpretable and explainable sentiment polarity detection for 400,000 concepts.

In our work, we choose SenticNet 6 and SenticNet 7, the two largest and latest sentiment knowledge bases, as our sources of external explicit sentiment information for words.

3. Methodology

3.1. Task definition

The MSA task aims to predict sentiment polarity or identify sarcasm, for a given video utterance. The multimodal dataset consists of N labeled utterances, $D = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_N\}$. Each utterance $\mathbf{U}_i$ consists of three modalities: text ($\mathbf{U}_i^t \in \mathbb{R}^{T_t}$), visual ($\mathbf{U}_i^v \in \mathbb{R}^{T_v \times d_v}$), and acoustic ($\mathbf{U}_i^a \in \mathbb{R}^{T_a \times d_a}$), respectively. Here $T_{\{t,v,a\}}$ and $d_{\{v,a\}}$ denote the sequence length and feature dimensions of the input modality. The corresponding labels for D are represented as $y = \{y_1, y_2, \dots, y_N\}$. The goal of our model is to fit the sentiment label y_i for video utterance $\mathbf{U}_i$.

3.2. Overall architecture

Our proposed approach can be segmented into five submodules: (1) the modality encoder (ME) module projects the text sequence and original visual/acoustic features into unimodal sentiment representation, (2) the TGI module interacts with the sentiment information through the interaction between text and non-verbal modalities, (3) the EKE module introduces external sentiment knowledge to further enhance multimodal sentiment fusion, (4) the FWAF module fuses the multi-source representations, (5) the prediction module to generate sentiment predictions. The overall architecture is illustrated in Fig. 1.

3.3. Modality encoder

Each input modality is first encoded into unimodal representations h_t, h_v, and h_a. For text modality, given a sequence of words $\mathbf{U}_i^t = \{x_1, x_2, \dots, x_{T_t}\}$, the pre-trained RoBERTa model is performed to get the token embeddings $\mathbf{H}_t$:

$$\mathbf{H}_t = RoBERTa(\mathbf{U}_i^t) \in \mathbb{R}^{T_t \times d}, \tag{1}$$

where d denotes the unified representation dimension of each modality. Then, the average pooling operation on $\mathbf{H}_t$ is used to obtain the text sentiment representation:

$$h_t = MeanPooling(\mathbf{H}_t), \tag{2}$$

where the average operation is performed on the token scale for each representation dimension.

As for visual and acoustic modalities, following previous works [6,18,22], a stack of unidirectional LSTM [55] coupled with a fully connected layer is then used to capture visual and acoustic sentiment embedding $\mathbf{H}_v$ and $\mathbf{H}_a$:

$$\mathbf{H}_m = FC(sLSTM(\mathbf{U}_i^m)) \in \mathbb{R}^{T_m \times d}, m \in \{v, a\}. \tag{3}$$

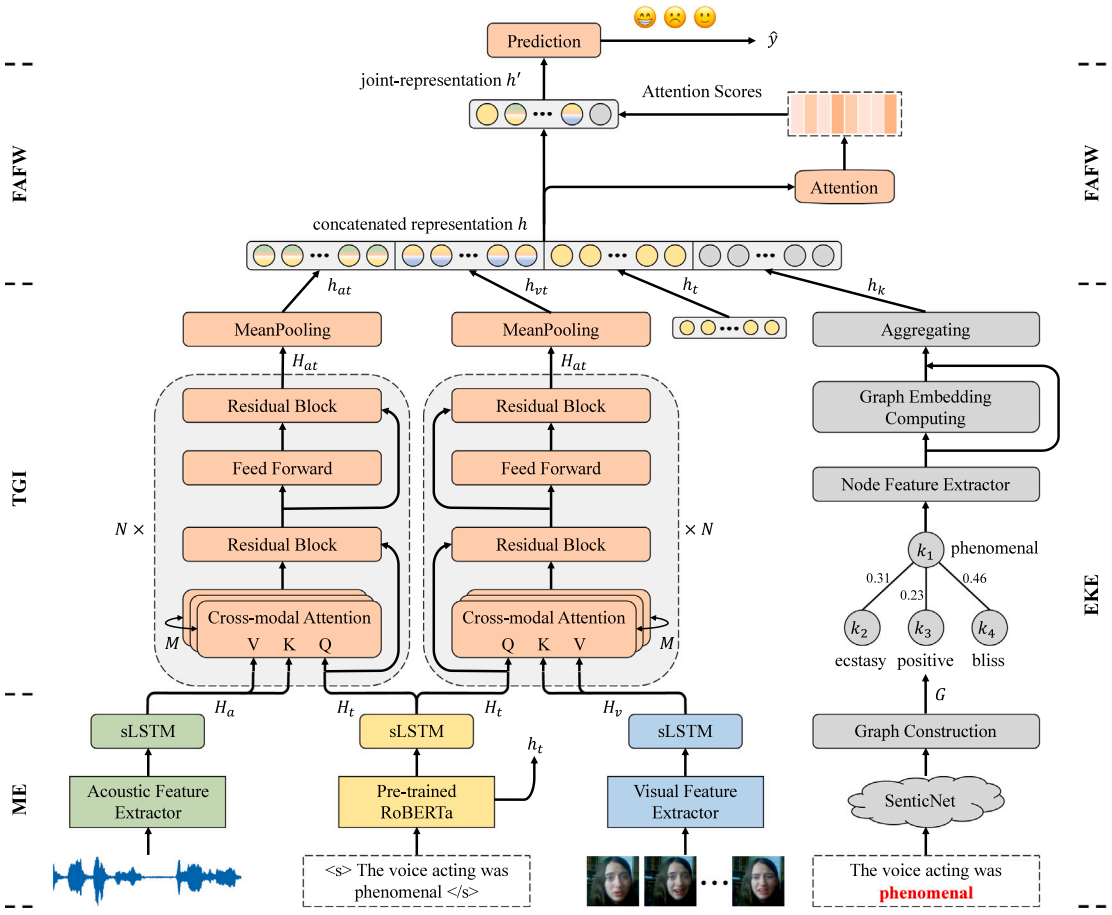


Fig. 1. The overall architecture of our proposed model for multimodal sentiment analysis. It consists of five components: ME for unimodal sentiment representation learning; TGI for text-guided interaction learning; EKE for external sentiment knowledge enhancement; FAWF for multimodal fusion; The last for sentiment prediction. The input of the model includes text, images, and audio sequences.

3.4. Text-guided interaction

After projecting the input modalities into unimodal representations, the TGI module would generate the crossmodal representation between text and visual ($\mathbf{H}_t$ and $\mathbf{H}_v$), text and acoustic ($\mathbf{H}_t$ and $\mathbf{H}_a$). This module considers the interaction between token pairs of text and non-verbal representation sequences. The core unit of the TGI module is crossmodal attention (CA), which is designed with reference to the self-attention mechanism [24]. The formula is as follows:

$$CA(\mathbf{H}_t, \mathbf{H}_m) = S\left(\frac{[\mathbf{H}_t\mathbf{W}^Q][\mathbf{H}_m\mathbf{W}^K]^T}{\sqrt{d}}\right)[\mathbf{H}_m\mathbf{W}^V], \tag{4}$$

where S is the Softmax function, $\mathbf{W}^{Q/K/V} \in \mathbb{R}^d$, $m \in \{v, a\}$. $\mathbf{W}^{Q/K/V}$ denotes the learned parameter matrices for the linear projections of query/key/value. In particular, for each layer, the text representation $\mathbf{H}_t$ is used as Query and non-verbal (visual or acoustic) representations $\mathbf{H}_v$ and $\mathbf{H}_a$ are used as Key and Value. In this way, the text representation can guide our model to pay more attention to relevant regions of non-verbal modalities.

To speed up training, this module performs M parallel attention functions with $d_M = d/M$ dimensional keys, values, and queries, and each attention function is called a head. The i_{th} head is computed as follows:

$$hd_i = CA_i(\mathbf{H}_t, \mathbf{H}_m). \tag{5}$$

The outputs of M heads are then concatenated and projected back to the original representation space, resulting in the final attention values:

$$MultiHead(\mathbf{H}_t, \mathbf{H}_m) = [hd_1 \oplus hd_2 \oplus \cdots \oplus hd_M]\mathbf{W}^O, \tag{6}$$

where $\mathbf{W}^O \in \mathbb{R}^{d\times d}$.

Following the self-attention mechanism [24], a residual block is used to avoid the degradation of the unimodal representation, which is depicted as:

$$\mathbf{H}_{mt} = LayerNorm(\mathbf{H}_t + MultiHead(\mathbf{H}_t, \mathbf{H}_m)), \tag{7}$$

where LayerNorm denotes the layer normalization that uses statistics computed from input data in both the training and evaluation process, and it is trainable.

After that, a fully connected layer and another residual block are utilized to get the output sequence of the crossmodal representation:

$$\mathbf{H}_{mt} = LayerNorm(\mathbf{H}_{tm} + FC(\mathbf{H}_t)) \in \mathbb{R}^{T_t\times d}. \tag{8}$$

Like the text modality encoder, an average operation is performed on the time scale for each feature dimension of $\mathbf{H}_{mt}$. Finally, the visual and acoustic sentiment representation h_{vt} and h_{at} guided by the text modality is obtained, where $h_{vt}/h_{at} \in \mathbb{R}^d$.

3.5. External knowledge enhancement

The EKE module aims to enhance the fusion embedding by generating additional sentiment knowledge representation. It is based on the external SenticNet knowledge base and network embedding methods. Specifically, a sentiment knowledge graph is first constructed for the text modality of each sample, and a graph embedding computing method is then applied to obtain the final sentiment knowledge enhancement embedding.

Graph Definition. Inspired by previous research on graph neural network [56,57], we represent a sentiment knowledge graph as $G = (V, E, H)$ where V and E is the set of the sentiment nodes and edges, respectively, and H denotes the set of features for each sentiment node. $e_{ij} = (v_i, v_j) \in E$ represents the edge across nodes v_i and v_j in V. To better express the relationship between sentiment nodes, $N(v)$ is also introduced to denote the neighborhood of node v, where $N(v) = \{u \in V | e_{vu} \in E\}$. The adjacency matrix A is a $n \times n$ matrix with n being the number of nodes.

Graph Construction. Since not all words from the text modality are qualified for SenticNet, the original text is filtered, and only the words with letters and numbers are reserved. Besides, it is worth noting that only a few words appear in the SenticNet for each sentence. Therefore, there are two types of nodes in the graph, V_1 is the words of the original text, and V_2 is the words' external sensitivity and emotional knowledge from SenticNet, $V_1 \cup V_2 = V$. The corresponding sentiment word embedding is also set as the node features H, where h_i for node v_i is extracted by the bert-base-uncased pre-trained model.

To build edges for the sentiment knowledge graph, we assume an edge e_{ij} exists between node v_i and v_j, where $v_i \in V_1$ and $v_j \in V_2$. To maximize the utilization of external sentiment knowledge, the mutual effect of internal nodes in V_1 and V_2 is ignored. Thus, the adjacency matrix A is defined as follows:

$$A_{ij} = \begin{cases} W_{ij} & v_i \in V_1, v_j \in V_2 \\ 0 & otherwise \end{cases}, \tag{9}$$

where W_{ij} is the weight of e_{ij} and is proportional to the number of co-occurrences of v_i and v_j. The frequency of each node v_i is further obtained:

$$P(v_i) = \frac{\sum_{j=0}^{n} A_{ij}}{\sum_{i=0}^{n}\sum_{j=0}^{n} A_{ij}}. \tag{10}$$

Finally, the positive pointwise mutual information (PPMI) matrix is computed to capture sentiment nodes' co-occurrence information:

$$PPMI(v_i, v_j) = \max(\log \frac{P(v_i, v_j)}{P(v_i)P(v_j)}, 0), \tag{11}$$

where $P(v_i, v_j)$ is the frequency that node v_i and v_j co-occur in sampled random walks. In our model, the $PPMI$ matrix is used to init the weight of the edges. The sentiment knowledge graph G is constructed with the above initialization of nodes and edges.

Graph Embedding Computing. Because the graph-level sentiment knowledge representation depends on each node, the sentiment feature for each node is first updated. For v_i in G, a binary multiplication operation between the node feature and its edge weights is then performed to obtain the information from neighbor nodes $N(v_i)$:

$$h_i' = \max_{j \in N(v_i)} h_i * PPMI(v_i, v_j), \tag{12}$$

where h_i is the node sentiment feature after message passing and aggregation.

Then the final sentiment feature of node v_i is computed as follows:

$$h_i = \alpha h_i + (1 - \alpha) h_i', \tag{13}$$

where α is a learnable parameter with a default value of 0, and α is used to dynamically adjust the weights and prevent the degradation of the node sentiment feature.

After aggregating node sentiment features, the graph-level sentiment knowledge embedding h_k is finally generated:

$$h_k = \mathcal{F}(h_1, h_2, \ldots, h_i, \ldots, h_n), \tag{14}$$

where $\mathcal{F}$ is a readout function for G, and a global average operation of each node is adopted here.

3.6. Feature-wised attention fusion

The FWAF module is designed to dynamically adjust the weight of each feature dimension of each modality and perform the multimodal fusion. The pipeline of this module is as follows: First, the four modality representations obtained above are concatenated into a multimodal representation: $h = [h_t \oplus h_{vt} \oplus h_{at} \oplus h_k]$, $h \in \mathbb{R}^{4\times d}$. Then, the weighted feature-wised attention matrix $\mathbf{W}_{Att}$ is computed:

$$\mathbf{W}_{Att} = \mathbf{W}_2(ReLU(\mathbf{W}_1 * h)), \tag{15}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable parameters, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{4\times d}$. In addition, the Sigmoid activation function ensures that each value in $\mathbf{W}_{Att}$ ranges between 0 and 1. In this way, the importance of each feature channel is acquired, which is then applied to enhance critical features' contribution to the multimodal fusion process. Finally, the concatenated representation is multiplied with $\mathbf{W}_{Att}$ to get the joint-representation h':

$$h' = \mathbf{W}_{Att} * h, \tag{16}$$

where h' is the final output of the FWAF module.

3.7. Prediction

In this part, a fully connected layer with a Softmax function is used to generate the sentiment prediction $\hat{y}_i$ for each utterance $\mathbf{U}_i$:

$$\hat{y}_i = Softmax(FC(h')). \tag{17}$$

Note that the layer normalization and dropout strategy are performed on h, enhancing the generalization of our model.

3.8. Optimization objectives

The L1Loss is used as the optimization objective on CMU-MOSI and CMU-MOSEI.

$$L = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| + \lambda \|\mathbf{W}\|_2^2, \tag{18}$$

where L is the loss, $\hat{y}_i$ is the regression result of our model for sample i, and y_i is the true value of sample i. N is the size of the training set. λ is the weight of the L2 regularization.

For Twitter2019, as the positive and negative samples are seriously unbalanced, the asymmetric loss (ASL) function [58] is adopted, which is designed for long-tailed distribution data:

$$L = -\frac{1}{N}\sum_{i=1}^{N} [y_i L_+ + (1 - y_i) L_-] + \lambda \|\mathbf{W}\|_2^2. \tag{19}$$

The L_+ and L_- are calculated as follows:

$$\begin{cases} L_+ = (1 - \hat{y}_i)^{\gamma_+} \log(\hat{y}_i) \\ L_- = \hat{y}_{im}^{\gamma_-} \log(1 - \hat{y}_{im}) \\ y_{im} = \max(y_i - m, 0) \end{cases}, \tag{20}$$

where γ_+ and γ_- are the focusing parameters for positive and negative data, and m is a hyper-parameter for tuning negative sample loss. In our experiments, we set 0 for γ_+, 4 for γ_-, and 0.1 for m.

4. Experiments

4.1. Datasets

We evaluated our model on two subtasks of MSA: multimodal sentiment intensity prediction, and multimodal sarcasm detection, with three datasets involved. The split specifications are shown in Table 1. Here, we give a brief introduction to the above datasets.

CMU-MOSI. The CMU-MOSI dataset [27] is one of the most popular benchmarks for the MSA task. It contains 93 YouTube video blogs

Table 1
Dataset statistics in CMU-MOSI, CMU-MOSEI, and Twitter2019. HG: Highly Negative, NG: Negative, WN: Weakly Negative, NU: Neutral, WP: Weakly Positive, PS: Positive, HP: Highly Positive, SA: Sarcasm. It is worth noting that the regression interval for CMU-MOSI/CMU-MOSEI is subdivided into seven sentiment categories since the classification metrics are needed to measure model performance.

Dataset	Modality	Sentiment	Type	Total	Train	Valid	Test
CMU-MOSI	Text, Acoustic, Visual	HN, NG, WN, NU, WP, PS, HP	Regression	2199	1284	229	686
CMU-MOSEI	Text, Acoustic, Visual	HN, NG, WN, NU, WP, PS, HP	Regression	22 856	16 326	1871	4659
Twitter2019	Text, Visual	SA, no SA	Classification	24 635	19 816	2410	2409

(vlogs) from 89 distinct English native speakers. The videos are sliced into 2199 short video clips, and each clip is manually annotated with a sentiment value ranging from −3 to 3, indicating the relative strength of expressed sentiment.

CMU-MOSEI. The CMU-MOSEI dataset [28] is the next generation of CMU-MOSI. And it is also the largest video-level English dataset of the MSA task. It comprises 23 453 annotated YouTube clips from 3228 videos and 1000 distinct speakers, involving 250 frequently used topics in online videos. The annotation strategy for this dataset is the same as CMU-MOSI.

Twitter2019. The Twitter2019 dataset [7] is a multimodal sarcasm detection dataset based on Twitter. Each tweet contains text and its corresponding picture, with 24 635 tweets. All the mentions in tweets are replaced with a certain symbol <user>. The label of each tweet is first annotated based on the hashtag that comes with the text (e.g., #sarcastic, etc.) and then checked manually.

4.2. Evaluation metrics

Multimodal sentiment intensity prediction. Following the previous works [5,14,18,22], we report our experimental results on regression and classification forms. For regression, mean absolute error (MAE) is the average absolute difference value between truth values and predicted values, and Pearson correlation (Corr) calculates the degree of prediction skew. For classification, we report seven-class classification accuracy (Acc-7) to indicate the proportion of predictions that are within the correct interval of seven intervals from strong negative (−3) to strong positive (3), binary classification accuracy (Acc-2) and weighted F1 score to measure non-negative/negative and positive/negative results. In all metrics except MAE, higher values indicate better performance.

Multimodal sarcasm detection. For sarcasm detection, following the previous works [7,40,41,59], we record precision (P) and recall (R) in addition to F1 and Acc.

4.3. Feature extraction

In order to ensure fair competition with the state-of-the-art, following previous works, we apply the following the data processing method for the three modalities below.

Text Features. All samples for the three datasets contains manual transcription, we fine-tune the roberta-base pre-trained model to acquire the text features. These features are 768-dimensional token embeddings.

Visual Features. The visual modality of CMU-MOSI and CMU-MOSEI is mainly contributed by faces, so the analytical tool called Facet[1] is used to extract facial expression features, including facial action units and face pose features. For Twitter2019, to extract the sequence features of the tweet images, we chop the last fully-connected (FC) and average pool layer of the ResNet-50 pre-trained model [60]. The sequence length of the features is fixed as 49, which is the same as the number of regions (7×7) of each input image. The visual feature dimensions are 47 for CMU-MOSI, 35 for CMU-MOSEI, and 2048 for Twitter2019.

Acoustic/OCR Features. For CMU-MOSI and CMU-MOSEI, a professional acoustic analysis framework named COVAREP [61] is used to extract 74-dimensional acoustic features. For Twitter2019, we use the popular optical character recognition (OCR) library PaddleOCR[2] to detect sentences in the image, and then utilize the bert-base-uncased pre-trained model to obtain OCR features. These BERT features are 768-dimensional token embeddings.

4.4. Baselines

To fully validate the performance of our model, we compared our results with the following state-of-the-art in the MSA task.

TFN [14]: This network modeling intra-modality (unimodal) and inter-modality (bimodal and trimodal) dynamics based on the vector field using the three-fold Cartesian product. It is an end-to-end model.

LMF [13]: This method utilizes low-rank tensors to avoid suffering from the exponential increase in dimensions, which can reduce the computational complexity and improve multimodal fusion efficiency.

MFM [37]: This model factorizes representations into two sets of independent factors. Multimodal discriminative factors are shared across all modalities, while Modality-Specific generative factors are unique for each modality. The joint generative-discriminative objective across multiple modalities and labels is the optimization goal of this model.

ICCN [33]: Compared with non-text features based on human-engineered, text features based on deep models typically have higher quality. This network learns correlations between multiple modalities via deep canonical correlation analysis (DCCA) and finally obtains the multimodal fusion embedding.

MulT [23]: The core of this method is directional crossmodal attention, which focuses on the interaction between multimodal sequences data (non-aligned data) across distinct time steps. It can also learn the long-range dependencies between elements of multiple modalities.

MISA [6]: The framework projects each modality to two distinct subspaces. Modality-Invariant learns the shared information and reduces the gap across modalities. Modality-Specific captures the characteristic information for each modality. All of the representations jointly provide an overall view of the multimodal data.

MAG-BERT [49]: This model includes a MAG attachment to allow multimodal nonverbal data during fine-tuning for pre-trained models like BERT. It generates a shift embedding to the language modality representation of BERT.

BBFN [18]: This approach performs the fusion process on pairwise modality representations based on an end-to-end architecture. It can integrate relevant information while reserving independent information from multiple modalities.

Self-MM [22]: Based on the unimodal supervision obtained by the self-supervised learning strategy, it takes the way of joint training the multimodal and unimodal tasks to learn the consistency and difference of modalities.

MMIM [5]: This model maintains critical task-related information by hierarchically maximizing the Mutual Information (MI). The MI is computed in unimodal input pairs (inter-modality) and between multimodal fusion embedding and unimodal inputs.

[1] https://imotions.com/platform/.

[2] https://github.com/PaddlePaddle/PaddleOCR.

Table 2
Hyper-parameters of SKEAFN for the multimodal sentiment analysis. For each dataset, the value on the left means the parameter with text encoder based on the pre-trained BERT, while the right one with text encoder based on the pre-trained RoBERTa.

Para	CMU-MOSI	CMU-MOSEI	Twitter2019
bs	64	128	64
vlr	3e−4/7e−4	7e−4/2e−4	6e−4/5e−4
alr	3e−3/2e−3	5e−3/6e−3	4e−3/1e−2
klr	9e−3/1e−2	1e−3/2e−3	2e−3/3e−3
flr	1e−3/8e−3	1e−2/1e−2	1e−3/1e−2
vwd	1e−3/5e−3	1e−3/4e−3	5e−3/4e−3
awd	4e−3/1e−2	7e−3/6e−3	2e−3/1e−2
kwd	1e−2/5e−3	5e−3/9e−3	2e−3/6e−3
fwd	1e−3/2e−3	4e−3/2e−3	4e−3/2e−3
N	2/5	3/2	2/2

Different from multimodal sentiment intensity prediction, the Twitter2019 dataset we used for validating the performance on the multimodal sarcasm detection task does not involve acoustic modality. The baselines are as follows.

HFM [7]: This model regards text features, image features, and image attributes as three modalities and carries multimodal fusion by leveraging bidirectional LSTM architecture. It is the first deep fusion model in multimodal sarcasm detection instead of simply concatenation.

D&R Net [41]: This method models cross-modality contrast and semantic association based on two networks. The decomposition network represents the commonality and difference between modalities, while the relation network builds semantic association of the context.

IIMI-MMSD [40]: This BERT architecture-based model uses inter-modality attention to capture inter-modality incongruity. In addition, it also applies the co-attention mechanism to model the contradiction within the text.

4.5. Experimental details

All models are implemented by PyTorch, running on a single NVIDIA Tesla V100-SXM2 GPU with 32G graphical memory size. The pre-trained BERT and RoBERTa model is available from the Transformer toolkit[3] released by Hugging Face. We fix the learning rate and weight decay for the pre-trained models as 5e−5 and 1e−3, respectively. We set different learning rate values and weight decay values for each rest module in SKEAFN, as shown in Table 2. We use Adam as the optimizer and take CosineAnnealingWarmRestarts as the scheduler for the learning rate. For the ME module, the number of sLSTM layers is set to 1. For the TGI module, the dimensionality of hidden representations is set to d = 768. The number of attention heads is fixed at M = 12. We set a different number of text-guided interaction layers N for each dataset. The dropout rate of 0.1 is utilized to avoid overfitting. Moreover, the early stopping with patience of 5 is used to get the best generalization performance.

5. Results and analysis

5.1. Comparison with baselines

The overall results of our proposed model and the baselines are reported in Tables 3, 4, and 5. The results demonstrate that the SKEAFN achieves significant improvement on all datasets we used. For a relatively fair comparison, we report results for both the SKEAFN with BERT text encoder $SKEAFN_1$ and the SKEAFN with RoBERTa text encoder $SKEAFN_2$. In the subsequent analysis, Acc-2 and F1 are "negative/positive" results.

[3] https://huggingface.co.

Table 3
Results on the test set of CMU-MOSI. The $SKEAFN_1$ uses BERT as text encoder, and $SKEAFN_2$ uses RoBERTa as text encoder. For Acc-2 and F1, the value on the left means the results of "negative/non-negative", while the right one is the results of "negative/positive".

Models	MAE(↓)	Corr (↑)	Acc-7 (↑)	Acc-2(↑)	F1 (↑)
TFN[a]	0.901	0.698	34.9	–/80.8	–/80.7
LMF[a]	0.917	0.695	33.2	–/82.5	–/82.4
MFM[a]	0.877	0.706	35.4	–/81.7	–/81.6
ICCN[a]	0.862	0.714	39.0	–/83.0	–/83.0
BBFN[b]	0.776	0.755	45.0	–/84.3	–/84.3
MulT[a]	0.861	0.711	–	81.5/84.1	80.6/83.9
MISA[a]	0.804	0.764	–	80.79/82.10	80.77/82.03
MAG-BERT[a]	0.727	0.781	43.62	82.37/84.43	82.50/84.61
Self-MM[c]	0.737	0.779	44.17	82.80/84.60	82.67/84.54
MMIM[c]	0.755	0.773	45.34	83.67/85.37	83.62/85.37
$SKEAFN_1$	0.740	0.784	45.18	84.40/86.43	84.50/86.47
$SKEAFN_2$	**0.665**	**0.825**	**47.08**	**85.13/87.34**	**85.18/87.34**

[a]Models are from [5].
[b]Models are from [18].
[c]Models are reproduced from open-source code with hyper-parameters provided in the original papers.

Table 4
Results on the test set of CMU-MOSEI. The $SKEAFN_1$ uses BERT as text encoder, and $SKEAFN_2$ uses RoBERTa as text encoder. For Acc-2 and F1, the value on the left means the results of "negative/non-negative", while the right one is the results of "negative/positive".

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)
TFN[a]	0.593	0.700	50.2	–/82.5	–/82.1
LMF[a]	0.623	0.677	48.0	–/82.0	–/82.1
MFM[a]	0.568	0.717	51.3	–/84.4	–/84.3
ICCN[a]	0.565	0.713	51.6	–/84.2	–84.2
BBFN[b]	0.529	0.767	**54.8**	–/86.2	–/86.1
MulT[a]	0.580	0.703	–	–/82.5	–/82.3
MISA[a]	0.568	0.724	–	82.59/84.23	82.67/83.97
MAG-BERT[a]	0.543	0.755	52.67	82.51/84.82	82.77/84.71
Self-MM[c]	0.535	0.763	53.48	79.46/84.16	80.18/84.24
MMIM[c]	0.543	0.758	52.75	83.28/85.04	83.48/84.88
$SKEAFN_1$	0.540	0.763	52.77	83.66/86.00	83.44/86.13
$SKEAFN_2$	**0.517**	**0.788**	54.21	**84.33/87.07**	**84.09/87.19**

[a]Models are from [5].
[b]Models are from [18].
[c]Models are reproduced from open-source code with hyper-parameters provided in the original papers.

Table 5
Results on the test set of Twitter2019. The $SKEAFN_1$ uses BERT as text encoder, and $SKEAFN_2$ uses RoBERTa as text encoder.

Models	Acc (↑)	P (↑)	R (↑)	F1 (↑)
HFM[a]	83.44	76.57	84.15	80.18
D&R Net[a]	84.02	77.97	83.42	80.60
IIMI-MMSD[a]	86.05	80.87	85.08	82.92
$SKEAFN_1$	85.76	81.62	82.89	82.25
$SKEAFN_2$	**93.77**	**90.89**	**93.74**	**92.29**

[a]The results for models are from [59].

Multimodal sentiment intensity prediction. Tables 3 and 4 show the comparative results of MOSI and MOSEI. On CMU-MOSI, our model outperforms the baseline models in general. Notably, $SKEAFN_2$ demonstrates a substantial improvement across all metrics, achieving an increase of **1.28%/1.36%** in Acc-2/F1, compared to the state-of-the-art model MMIM [5]. On CMU-MOSEI, the $SKEAFN_2$ reaches the new state-of-the-art in all metrics except Acc-7. Moreover, our model attains an increase of **0.87%/1.09%** in Acc-2/F1. In addition, the performance of the $SKEAFN_2$ is significantly better than the $SKEAFN_1$, with an increase of **0.91%/1.07%** in Acc-2 on CMU-MOSI/CMU-MOSEI. Nonetheless, the $SKEAFN_1$ still achieves comparable or even better results with

Table 6
Results of ablation study for SKEAFN's each modality input on the test set of CMU-MOSI. For Acc-2 and F1, the value on the left means the results of "negative/non-negative", while the right one is the results of "negative/positive".

Num	Model	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)
1	Text	**0.687**	**0.812**	**45.48**	**83.96/85.82**	**83.98/85.78**
2	Visual	1.403	0.060	15.45	55.24/57.77	71.17/73.23
3	Acoustic	1.443	0.081	15.45	44.75/42.25	61.83/59.37
4	External Knowledge	1.138	0.518	25.94	68.51/70.73	68.59/70.66
5	Text, Visual	**0.680**	**0.820**	**50.14**	83.67/86.12	83.73/86.11
6	Text, Acoustic	0.780	0.801	43.14	83.38/85.21	83.35/85.14
7	Visual, Acoustic	1.393	0.068	15.45	58.23/55.97	59.80/61.57
8	Text, External Knowledge	0.685	0.818	45.48	**84.11/86.73**	**84.20/86.75**
9	Visual, External Knowledge	1.098	0.517	26.09	69.97/71.95	70.20/72.04
10	Acoustic, External Knowledge	1.104	0.529	27.84	69.53/71.64	69.86/71.81
11	Text, Visual, Acoustic	0.702	0.824	44.02	84.83/86.43	84.87/86.42
12	Text, Visual, External Knowledge	**0.668**	**0.825**	46.50	**84.83/87.19**	**84.92/87.21**
13	Text, Acoustic, External Knowledge	0.689	0.817	**46.64**	84.54/86.43	84.54/86.38
14	Visual, Acoustic, External Knowledge	1.128	0.515	27.25	70.69/72.86	70.67/72.72
15	Text, Visual, Acoustic, External Knowledge	**0.665**	**0.825**	**47.08**	**85.13/87.34**	**85.18/87.34**

Table 7
Results of ablation study for SKEAFN's functional submodules on the test set of CMU-MOSI. "w/o" denotes removing some submodules from the whole SKEAFN. For Acc-2 and F1, the value on the left means the results of "negative/non-negative", while the right one is the results of "negative/positive".

Model	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)
SKEAFN	**0.665**	**0.825**	**47.08**	**85.13/87.34**	**85.18 /87.34**
w/o TGI	0.698 (↑ 0.033)	0.798 (↓ 0.027)	46.06 (↓ 1.02)	84.25 (↓ 0.88)/86.43 (↓ 0.91)	84.87 (↓ 0.31)/86.42 (↓ 0.92)
w/o EKE	0.702 (↑ 0.037)	0.824 (↓ 0.001)	44.02 (↓ 3.06)	84.83 (↓ 0.30)/86.43 (↓ 0.91)	84.87 (↓ 0.31)/86.42 (↓ 0.92)
w/o FWAF	0.701 (↑ 0.036)	0.791 (↓ 0.034)	43.87 (↓ 3.21)	83.38 (↓ 1.75)/85.67 (↓ 1.67)	83.43 (↓ 1.75)/85.65 (↓ 1.69)
w/o TGI and FWAF	0.713 (↑ 0.048)	0.802 (↓ 0.023)	45.18 (↓ 1.90)	82.65 (↓ 2.48)/84.75 (↓ 2.59)	82.66 (↓ 2.52)/84.70 (↓ 2.64)
w/o EKE and FWAF	0.718 (↑ 0.053)	0.792 (↓ 0.033)	42.41 (↓ 4.67)	83.09 (↓ 2.04)/85.06 (↓ 2.28)	83.15 (↓ 2.03)/85.06 (↓ 2.28)

the state-of-the-art in the same experimental environment, taking the increase of **1.06%/0.96%** in Acc-2 on CMU-MOSI/CMU-MOSEI as an example.

Multimodal sarcasm detection. The results on Twitter2019 are reported in Table 5. We can observe that $SKEAFN_2$ outperforms the baseline models, exhibiting a significant improvement of **7.72%** and **9.37%** in terms of Acc and F1, respectively. Conversely, $SKEAFN_1$ yields lower results than $SKEAFN_2$, showing a decrease of 8.01% and 10.04% in Acc and F1, respectively. This outcome can be attributed to the fact that RoBERTa provides better text representation compared to BERT. However, $SKEAFN_1$ achieves comparable and even superior results, with a **0.75%** increase in P compared to IIMI-MMSD [40], without introducing image objective detection.

5.2. Ablation study

5.2.1. Effect of each modality for SKEAFN

To further evaluate the contributions of each modality in SKEAFN, we conduct three sets of ablation experiments on CMU-MOSI, as shown in Table 6. In the following analysis, we mainly focus on the metric of Acc-2.

The first experiment set (uni-modal) aims to explore the individual role of each modality, corresponding to experiments numbered 1–4 in Table 6. We only use one modality as input, while the others are excluded. The results indicate that: (1) Among the four modalities, the text modality achieves the best results, with results lagging only 1.52% compared to the SKEAFN. This shows the decisive role of text modality. (2) The kb modality's performance outperforms the visual and acoustic modalities by 12.96% and 28.48% respectively. It powerfully demonstrates the rationality of the introduction of external knowledge. To combine (1) and (2), we also find that the sentiment information contained in non-verbal modalities (visual and acoustic) is more difficult to recognize than in verbal modalities (text and kb).

The second experiment set (bi-modal) reveals the interaction between the two modalities. Six distinct bimodal combinations correspond to experiments numbered 5–10 in Table 6. From the results, we can make the following observations: (1) The performance of two modalities is generally better than that of a single modality. This preliminarily proves that there is complementary sentiment information in different modalities. (2) The introduction of external knowledge can promote the performance of a single modality (increased by 0.91% for text, 14.18% for visual, and 29.39% for acoustic). It indicates that external knowledge can enhance the sentiment semantic understanding ability of our model.

The third experiment set (tri-modal) further investigates the correlation among three modalities. According to the experiments numbered 11–14 in Table 6, the following observations can be listed: (1) The performance of the three modalities has further improved compared to the two modalities. This also proves the complementarity of the three modalities. (2) It has insignificant improved (less than 1%) from two to three modalities. Furthermore, the best result of the three modalities is only 0.15% less than the SKEAFN, which is very close. These two observations elaborate on the effectiveness of multiple modality combinations, and we can also conclude that either modality is indispensable.

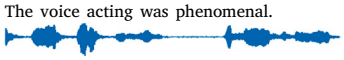
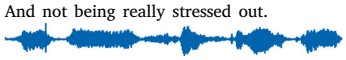
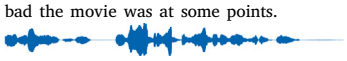
5.2.2. Effect of different submodules in SKEAFN

We conduct several experiments on CMU-MOSI to investigate the effects of different submodules in SKEAFN. We remove (1) TGI, (2) EKE, (3) FWAF, (4) TGI and FWAF, (5) EKE and FWAF, respectively. The results are reported on Table 7. We can observe that the result declines when removing either TGI, EKE, or FWAF in the SKEAFN. Especially, removing FWAF has the greatest impact on the metrics, with a **1.67%** decrease in Acc-2 and a **3.21%** decrease in Acc-7. Removing EKE leads a **0.91%** decrease in Acc-2 and a **3.06%** decrease in Acc-7. It proves that TGI, EKE, and FWAF are all critical for the SKEAFN. It is also clear that removing TGI and FWAF (or EKE and FWAF) simultaneously results in a significant drop. It indicates the effectiveness of the cooperation among submodules, which can benefit the multimodal fusion and promote sentiment prediction.

5.3. Case study

Table 8 presents some examples where our proposed model predicts according to four modalities. For case A, the word "phenomenal"

Table 8
Representative examples on the MOSI dataset with their predictions and truth scores in our case study. The Absolute Error is the absolute difference between model prediction and ground truth.

Case	Input modality description	Prediction	Ground truth	Absolute error
A	Frown + Closed eyes The voice acting was phenomenal. 	2.754	2.799	0.045
B	Tilted head + Wrinkled nose + Squinting And not being really stressed out. 	−0.023	0.000	0.023
C	Raised eyebrows + Opened mouth I was sitting there in the theatre laughing how bad the movie was at some points. 	−1.599	−1.600	0.001

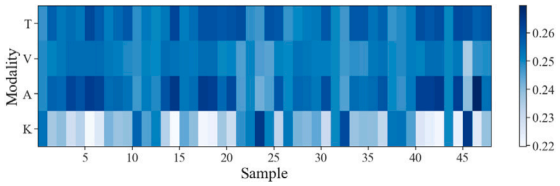


Fig. 2. Visualization of the feature-wised attention matrices for some samples on the test set of CMU-MOSI. Each column denotes an example, and each row shows the attention scores for the corresponding modality.

involves extreme positive sentiment tendencies and explicit positive sentiment in both text modality and KB modality, respectively. For case B, the word "really" provides additional explicit positive information in the kb modality, which finally obtains a neutral recognition mutually with negative information from visual and acoustic modalities. For case C, the word "bad" contributes a solid negative sentiment clue from the kb modality and enhances the negative polarity for the text modality while weakening the effect of the word "laughing". Therefore, our model attains a very close prediction compared to the ground truth. These examples demonstrate that with our model, both verbal and non-verbal modalities information, especially the explicit sentiment information provided by the kb modality, is helpful to the sentiment analysis.

5.4. Visualization

We take 48 samples from the CMU-MOSI test set and visualize the corresponding feature-wised attention scores for each modality's representation on these samples. As shown in Fig. 2, each column denotes the feature-wised attention scores for text, visual, acoustic, and external knowledge sentiment representation for a sample. The color scheme indicates the score intensity, with navy blue representing the highest score and light blue indicating the lowest. We can intuitively find that the importance of each modality in the multimodal fusion process is different. In general, the area of text and acoustic is darker than the area of visual and external knowledge, which means that text and acoustic have more influence on the generation of the joint-representation than visual and external knowledge. Nevertheless, we still find that the external knowledge representation plays a critical role in some samples corresponding to all samples with the darkest area in K row. It indicates that the FWAF module can mine the critical information from sentiment representations and assign them different weights according to their importance. It also indicates the necessity of introducing external knowledge, further demonstrating the effectiveness of the SKEAFN.

5.5. Discussion

In this subsection, we provide a brief summary of all the experimental results by systematically discussing the following two key questions.

Does the introduction of external sentiment knowledge benefit the MSA task? The comparison results demonstrate that our proposed SKEAFN model achieves superior performance in two subtasks of the MSA task. Specifically, the model achieves accuracies of 85.13%/87.34%, 84.33%/87.07%, and 93.77% on the CMU-MOSI, CMU-MOSEI, and Twitter2019 datasets, respectively, as presented in Tables 3, 4, and 5. These results illustrate that introducing external sentiment knowledge can lead to state-of-the-art performance. Similarly, the ablation studies in Section 5.2.1 indicate that excluding external sentiment knowledge would result in a noticeable degradation in all experiments. Coincidentally, the case study on three illustrative samples in Section 5.3 intuitively illustrates that leveraging external explicit sentiment information provided by the kb modality can effectively prompt the MSA model to achieve more accurate predictions. All these observations point to the fact that the introduction of external sentiment knowledge does benefit the MSA task.

Are the attention-based fusion strategies effective? In Section 5.2.1, three sets of ablation studies, examining uni-modal, bi-modal, and tri-modal configurations, confirm the indispensability of fusion strategies for the MSA task. These results also highlight the importance and effectiveness of the TGI module, which performs multimodal fusion using a text-guided attention mechanism. Additionally, the visualization results of the attention scores in Section 5.4 provide substantial evidence of the irreplaceable role played by the FWAF module in adjusting the contribution of each modality's representation. Furthermore, the ablation studies in Section 5.2.2 directly demonstrate the critical contribution of the TGI and FWAF modules in multimodal fusion to the final SKEAFN model. Overall, these observations strongly support the conclusion that the attention-based fusion strategies in the SKEAFN model are effective for the MSA task.

6. Conclusion

In this paper, we propose SKEAFN, a novel end-to-end fusion network that leverages external explainable sentiment knowledge to enhance multimodal fusion for the MSA task. The architecture of the SKEAFN model is mainly composed of the EKE module to generate additional sentiment knowledge representation, the TGI module to model the interactions between text and acoustic/visual modalities, and the FWAF module to fuse multimodal information. The EKE module provides sentiment semantic prompt information that can be captured through graph embedding computing. Simultaneously, the TGI module

derives text-guided crossmodal sentiment representations using stacked multi-head attention layers. Lastly, the FWAF module dynamically adjusts the weights of the additional sentiment knowledge representation and each modality's representations using feature-wised attention. Comprehensive experiments demonstrate that SKEAFN outperforms other baseline approaches and achieves state-of-the-art performance. In addition, the effectiveness of both the strategies of introducing external explicit sentiment knowledge and attention-based fusion is also validated by ablation study, case study, and visualization analysis.

In the future, we plan to conduct more experiments on other multimodal sarcasm detection datasets to establish new benchmarks in the field of MSA. We also intend to explore additional multimodal fusion methods based on external sentiment knowledge, aiming to enhance model interpretability. Furthermore, we will leverage massive unlabeled sentiment data to improve model performance by using semi-supervised and unsupervised learning methods.

CRediT authorship contribution statement

Chuanbo Zhu: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Min Chen:** Conceptualization, Writing – review & editing. **Sheng Zhang:** Methodology, Writing – review & editing. **Chao Sun:** Writing – original draft. **Han Liang:** Writing – original draft. **Yifan Liu:** Writing – original draft. **Jincai Chen:** Writing – review & editing, Funding acquisition.

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 link to my code at the revised manuscript

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62272178.

References

[1] 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, http://dx.doi.org/10.1109/JSTSP.2020.2987728.

[2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[3] 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, http://dx.doi.org/10.1016/j.knosys.2021.107676.

[4] R. Wang, Y. Hao, Q. Yu, M. Chen, I. Humar, G. Fortino, Depression analysis and recognition based on functional near-infrared spectroscopy, IEEE J. Biomed. Health Inf. 25 (12) (2021) 4289–4299, http://dx.doi.org/10.1109/JBHI.2021.3076762.

[5] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Dominican Republic, 2021, pp. 9180–9192, http://dx.doi.org/10.18653/v1/2021.emnlp-main.723, Online and Punta Cana, URL: https://aclanthology.org/2021.emnlp-main.723.

[6] 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, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1122–1131, http://dx.doi.org/10.1145/3394171.3413678.

[7] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in Twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2506–2515, http://dx.doi.org/10.18653/v1/P19-1239.

[8] J. Wang, L. Sun, Y. Liu, M. Shao, Z. Zheng, Multimodal sarcasm target identification in tweets, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8164–8175, http://dx.doi.org/10.18653/v1/2022.acl-long.562, URL: https://aclanthology.org/2022.acl-long.562.

[9] M.S. Akhtar, D.S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis, 2019, arXiv preprint arXiv:1905.05812.

[10] D.S. Chauhan, D. S R, A. Ekbal, P. Bhattacharyya, Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4351–4360, http://dx.doi.org/10.18653/v1/2020.acl-main.401, Online, URL: https://aclanthology.org/2020.acl-main.401.

[11] M. Chen, K. Shen, R. Wang, Y. Miao, Y. Jiang, K. Hwang, Y. Hao, G. Tao, L. Hu, Z. Liu, Negative information measurement at AI edge: A new perspective for mental health monitoring, ACM Trans. Internet Technol. 22 (3) (2022) http://dx.doi.org/10.1145/3471902.

[12] T. Jin, S. Huang, Y. Li, Z. Zhang, Dual low-rank multimodal fusion, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 377–387, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.35, Online.

[13] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Bagher Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2247–2256, http://dx.doi.org/10.18653/v1/P18-1209.

[14] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114, http://dx.doi.org/10.18653/v1/D17-1115.

[15] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, Proc. AAAI Conf. Artif. Intell. 34 (01) (2020) 164–172, http://dx.doi.org/10.1609/aaai.v34i01.5347.

[16] S. Mai, S. Xing, J. He, Y. Zeng, H. Hu, Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling, ACM Trans. Multimedia Comput. Commun. Appl. (2022) http://dx.doi.org/10.1145/3542927, Just Accepted.

[17] L. Xiao, X. Wu, W. Wu, J. Yang, L. He, Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 4578–4582, http://dx.doi.org/10.1109/ICASSP43922.2022.9747542.

[18] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 6–15, http://dx.doi.org/10.1145/3462244.3479919.

[19] K. Yang, H. Xu, K. Gao, CM-BERT: Cross-modal BERT for text-audio sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 521–528, http://dx.doi.org/10.1145/3394171.3413690.

[20] X. Zhao, Y. Chen, W. Li, L. Gao, B. Tang, MAG+: An extended multimodal adaptation gate for multimodal sentiment analysis, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 4753–4757, http://dx.doi.org/10.1109/ICASSP43922.2022.9746536.

[21] S. Zhang, M. Chen, J. Chen, Y.-F. Li, Y. Wu, M. Li, C. Zhu, Combining cross-modal knowledge transfer and semi-supervised learning for speech emotion recognition, Knowl.-Based Syst. 229 (2021) 107340, http://dx.doi.org/10.1016/j.knosys.2021.107340.

[22] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, Proc. AAAI Conf. Artif. Intell. 35 (12) (2021) 10790–10797.

[23] 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 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569, http://dx.doi.org/10.18653/v1/P19-1656.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[25] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 105–114, http://dx.doi.org/10.1145/3340531.3412003.

[26] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of LREC 2022, 2022.

[27] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88, http://dx.doi.org/10.1109/MIS.2016.94.

[28] A. Bagher 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246, http://dx.doi.org/10.18653/v1/P18-1208.

[29] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, CoRR abs/1810.04805, arXiv:1810.04805, URL: http://arxiv.org/abs/1810.04805.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR abs/1907.11692, arXiv:1907.11692, URL: http://arxiv.org/abs/1907.11692.

[31] A. Ghorbanali, M.K. Sohrabi, F. Yaghmaee, Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks, Inf. Process. Manage. 59 (3) (2022) 102929.

[32] M.U. Salur, İ. Aydın, A soft voting ensemble learning-based approach for multimodal sentiment analysis, Neural Comput. Appl. (2022) 1–16.

[33] Z. Sun, P.K. Sarma, W.A. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 8992–8999, URL: https://ojs.aaai.org/index.php/AAAI/article/view/6431.

[34] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 59–66.

[35] R. Chen, W. Zhou, Y. Li, H. Zhou, Video-based cross-modal auxiliary network for multimodal sentiment analysis, IEEE Trans. Circuits Syst. Video Technol. (2022) 1, http://dx.doi.org/10.1109/TCSVT.2022.3197420.

[36] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Košir, Audio-visual emotion fusion (AVEF): A deep efficient weighted approach, Inf. Fusion 46 (2019) 184–192, http://dx.doi.org/10.1016/j.inffus.2018.06.003.

[37] Y.-H.H. Tsai, P.P. Liang, A. Zadeh, L.-P. Morency, R. Salakhutdinov, Learning factorized multimodal representations, 2019, ArXiv abs/1806.06176.

[38] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, J. Qian, Multimodal sentiment analysis with image-text interaction network, IEEE Trans. Multimed. (2022).

[39] 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, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 2399–2402, http://dx.doi.org/10.1145/3132847.3133142.

[40] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1383–1392, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.124, Online.

[41] N. Xu, Z. Zeng, W. Mao, Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 3777–3786, http://dx.doi.org/10.18653/v1/2020.acl-main.349, Online, URL: https://aclanthology.org/2020.acl-main.349.

[42] W. Jin, B. Zhao, L. Zhang, C. Liu, H. Yu, Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis, Inf. Process. Manage. 60 (3) (2023) 103260, http://dx.doi.org/10.1016/j.ipm.2022.103260, URL: https://www.sciencedirect.com/science/article/pii/S0306457322003612.

[43] W. Jin, B. Zhao, H. Yu, X. Tao, R. Yin, G. Liu, Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning, Data Min. Knowl. Discov. 37 (1) (2023) 255–288, http://dx.doi.org/10.1007/s10618-022-00891-8, URL: https://link.springer.com/article/10.1007/s10618-022-00891-8.

[44] M. Hou, J. Tang, J. Zhang, W. Kong, Q. Zhao, Deep multimodal multilinear fusion with high-order polynomial pooling, Adv. Neural Inf. Process. Syst. 32 (2019).

[45] Q. Chen, G. Huang, Y. Wang, The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis, IEEE/ACM Trans. Audio, Speech, Lang. Process. (2022) 1–8, http://dx.doi.org/10.1109/TASLP.2022.3192728.

[46] 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 Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, in: AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[47] H. Zou, Y. Si, C. Chen, D. Rajan, E.S. Chng, Speech emotion recognition with co-attention based multi-level acoustic information, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7367–7371, http://dx.doi.org/10.1109/ICASSP43922.2022.9747095.

[48] 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 Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, in: AAAI'19/IAAI'19/EAAI'19, AAAI Press, 2019, http://dx.doi.org/10.1609/aaai.v33i01.33017216.

[49] W. Rahman, M.K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 2359–2369, http://dx.doi.org/10.18653/v1/2020.acl-main.214, Online.

[50] A. Esuli, F. Sebastiani, Sentiwordnet: A publicly available lexical resource for opinion mining, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), 2006.

[51] G.A. Miller, WordNet: A lexical database for english, Commun. ACM 38 (11) (1995) 39–41, http://dx.doi.org/10.1145/219717.219748.

[52] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010.

[53] E. Cambria, A. Hussain, SenticNet, in: Sentic Computing, Springer, 2015, pp. 23–71.

[54] E. Cambria, R. Speer, C. Havasi, A. Hussain, Senticnet: A publicly available semantic resource for opinion mining, in: 2010 AAAI Fall Symposium Series, 2010.

[55] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[56] H.L. Nguyen, D.T. Vu, J.J. Jung, Knowledge graph fusion for smart systems: A survey, Inf. Fusion 61 (2020) 56–70.

[57] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2020) 4–24.

[58] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 82–91, http://dx.doi.org/10.1109/ICCV48922.2021.00015.

[59] S. Gupta, A. Shah, M. Shah, L. Syiemlieh, C. Maurya, FiLMing multimodal sarcasm detection with attention, in: T. Mantoro, M. Lee, M.A. Ayu, K.W. Wong, A.N. Hidayanto (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2021, pp. 178–186.

[60] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[61] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP — A collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 960–964, http://dx.doi.org/10.1109/ICASSP.2014.6853739.