KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis

Anonymous Authors

ABSTRACT

Multimodal sentiment analysis (MSA) aims to integrate multiple modalities of information to better understand human sentiment. The current research mainly focuses on conducting multimodal fusion and representation learning, which neglects the under-optimized modal representations generated by the imbalance of unimodal performances in joint learning. Moreover, the size of labeled datasets limits the generalization ability of existing supervised models used in MSA. To address the above issues, this paper proposes a knowledge-enhanced self-supervised balanced representation approach (KEBR) to capture common sentimental knowledge in unlabeled videos and explore the optimization issue of information imbalance between modalities. First, a text-based cross-modal fusion method (TCMF) is constructed, which injects the non-verbal information from the videos into the semantic representation of text to enhance the multimodal representation of text. Then, a multimodal cosine constrained loss (MCC) is designed to constrain the fusion of non-verbal information in joint learning to balance the representation of multimodal information. Finally, with the help of sentiment knowledge and non-verbal information, KEBR conducts sentiment word masking and sentiment intensity prediction, so that the sentiment knowledge in the videos is embedded into the pre-trained multimodal representation in a balanced manner. Experimental results on two publicly available datasets MOSI and MOSEI show that KEBR significantly outperforms the baseline, achieving new state-of-the-art results.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Multimodal sentiment analysis, Knowledge enhanced pre-training, Imbalanced optimization, Text-based cross-modal fusion method, Multimodal cosine constrained loss

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) offers a comprehensive understanding of human sentiment by integrating information from text, audio, and visual modalities, which are closer to real-life scenarios where human beings process sentiment [2]. The widespread use of smart devices has led to a significant increase in user opinion videos, which offer abundant data resources and application scenarios for studying MSA tasks, such as depression detection, e-commerce, smart customer service, and human-computer, interaction, among others. With the generation of large amounts of multi-source information, MSA has received increasing attention, and richer information can also help to improve model performance [42]. However, due to the heterogeneous modal gap, achieving human-comparable MSA performance remains still challenging [44].

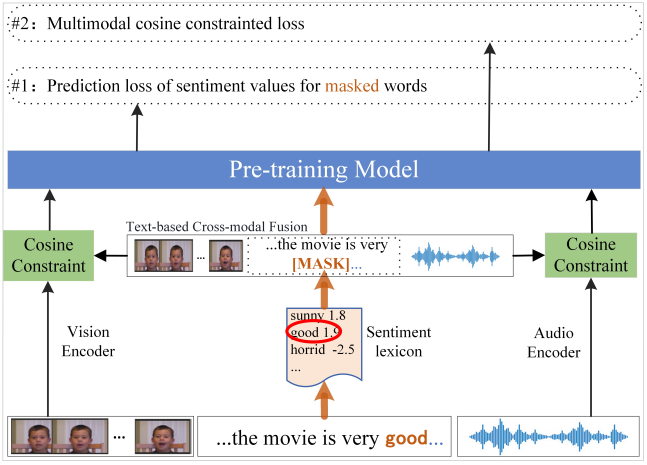

Figure 1: The pipeline of KEBR. The red ellipse represents the sentiment words and corresponding sentiment values in the text searched from the sentiment lexicon.

The majority of previous MSA approaches have concentrated on developing multimodal fusion and representation learning. RNN-based models[6, 20, 21, 51] connect each modal feature to a fusion vector input for subsequent classification or regression. Models based on Transformers [35, 36] are employed to simulate multimodal interactions for reducing the effect of inter-modal differences. The BERT [12, 29, 40, 45] pre-trained language model is employed as an encoder for text modality, thereby enhancing the performance of the MSA task by leveraging BERT's exceptional language representation capabilities. However, a joint model for uniform objective optimization may have its unimodal encoders converging at different rates [40]. One dominant modality could dominate the optimization process, leading to the neglect of other modalities and causing under-optimized representations. This results in modal bias that fails to fully utilize the capabilities of multimodal. To address this issue, some studies control the learning rate of different modalities [24, 26, 38]. However, all of these methods only mitigate the imbalance between the audio-visual modalities, and they all require additional training costs. Numerous studies have demonstrated

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

that text plays a more central role than non-verbal (audio, vision) information in MSA tasks[10, 22, 30, 31, 33, 37, 49]. To explore the dynamic relationship between modalities, HyCon [22] and ConKI [49] conducted intra-modal and inter-modal comparative learning to explore the interaction between modalities, but heavily relied on labeled datasets. Due to the limited availability of MSA sentiment-annotated datasets and the increased parameter capacity of fused models, supervised models tend to experience significant overfitting [8, 38], thus reducing the generalization ability of existing MSA models.

Through the above analysis, to improve the quality and performance of fusion representations for MSA, two key issues need to be addressed: 1) How to avoid modal imbalance (mainly non-verbal) due to modal gaps and efficiently construct text-centered non-verbal joint multimodal representations? 2) How to address the problem of overfitting and poor generalization ability in supervised models when using limited labeled datasets?

Inspired by knowledge-enhanced pre-training models on text sentiment analysis [32, 34, 54, 55], we found that a large number of unlabeled opinion videos on the Internet contain valuable sentiment knowledge. This knowledge can guide the fusion of visual, audio, and text modalities to express sentiment patterns or combine sentiment semantics. Learning this knowledge will enhance sentiment representation in MSA tasks. This can enhance further learning on a restricted MSA dataset. To this end, we propose a knowledge-enhanced self-supervised balanced representation (KEBR) method. Sentiment knowledge helps predict the sentiment intensity of masked sentiment words in unlabeled opinion videos using contextual and non-verbal information to learn common sentiment patterns. Specifically, we propose a text-based cross-modal fusion method (TCMF). Different from the previous work, we employ multi-layer cross-modal fusion to inject low-level features of non-verbal modalities into the semantic representation of text, while keeping the fused parameter capacity unchanged. This approach preserves the original affective semantics of the non-verbal information and enhances the text-based multimodal representation. In addition, a multimodal cosine constrained loss (MCC) was designed to mitigate the modal imbalance. Different from previous methods, MCC is capable of optimizing the imbalance among the three modalities. Its design restricts the injected non-verbal information in the fused multimodal representations to mitigate modal bias. Furthermore, MCC is designed as an external constraint with almost no additional training cost and is independent of the model or architecture.

As shown in Fig.1, first, the most significant sentiment word in the text is masked according to the pre-specified sentiment lexicon. Then, employing the proposed text-based cross-modal fusion method, the injection fusion of non-verbal information and the enhancement of text representation is accomplished. At the same time, the designed multimodal cosine constrained loss is applied to avoid the problem of under-optimization of non-verbal information due to modal gaps in the joint multimodal representation. Finally, the masked word representations are used to predict the joint loss of sentiment intensity and multimodal cosine constraints, embedding word-level sentiment knowledge from the video into pre-trained multimodal representations in a balanced manner.

After pre-training KEBR, to evaluate its effectiveness, we fine-tuned it on two benchmark datasets: MOSI [52] and MOSEI [1]. The experimental results demonstrate that KEBR outperforms the baseline and achieves state-of-the-art performance. The code has been released at https://github.com/*****/KEBR/.

The primary contributions of this paper are as follows:

- We propose a knowledge-enhanced multimodal self-supervised balanced representation method, which leverages the sentiment knowledge from large-scale unlabeled videos to enhance multimodal sentiment representation learning.
- We propose a text-based cross-modal fusion approach that injects low-level features from audio and visual modalities into text to enhance the multimodal information representation of text. This method highlights the dominance of text and the assistance of non-verbal information in modal fusion, which preserves the original affective semantics of the non-verbal information while enhancing the text-based multimodal representation.
- We propose a multimodal cosine constrained loss function to optimize the imbalance of unimodal in joint representation, which avoids the problem of certain modalities being neglected in fusion unable to exploit their capabilities.

2 RELATED WORK

2.1 Knowledge Enhanced Pre-training Language Models

In NLP, Transformer-based pre-trained language models have been widely studied and applied to extract contextual semantic features. Considerable attention has been focused on pre-training models on large-scale unlabeled datasets [9, 18, 46] to capture linguistic information, followed by fine-tuning the model for specific downstream tasks.

It is effective to introduce domain-specific knowledge into the process of pre-training language models [53]. The domain-specific knowledge can be the common knowledge of entity type and relationship classification tasks [17, 27, 53], legal knowledge of extracting legal elements [56], sentiment knowledge of sentiment analysis [15, 34, 47], and biomedical knowledge for health question-and-answer and medical reasoning [13]. Several studies [17, 27, 53, 54] integrate sentiment knowledge, encompassing sentiment words, word polarity, and aspect-emotion pairs, into text representations to improve specific sentiment analysis tasks. Knowledge-enhanced pre-trained language models have made significant progress in text sentiment analysis. However, few studies have utilized the abundant sentiment knowledge presented in unlabeled videos to improve sentiment representation learning for MSA tasks.

2.2 Multimodal Sentiment Analysis

MSA collects and processes data from multiple sources of audio, visual, and text information to comprehensively understand human sentiment [11]. Early MSA studies integrate multimodal representations obtained from different feature extraction networks [28]. The Tensor Fusion Network (TFN) [50] learns the intra-modaldynamics through modal embedding sub-networks. Low-rank multimodal

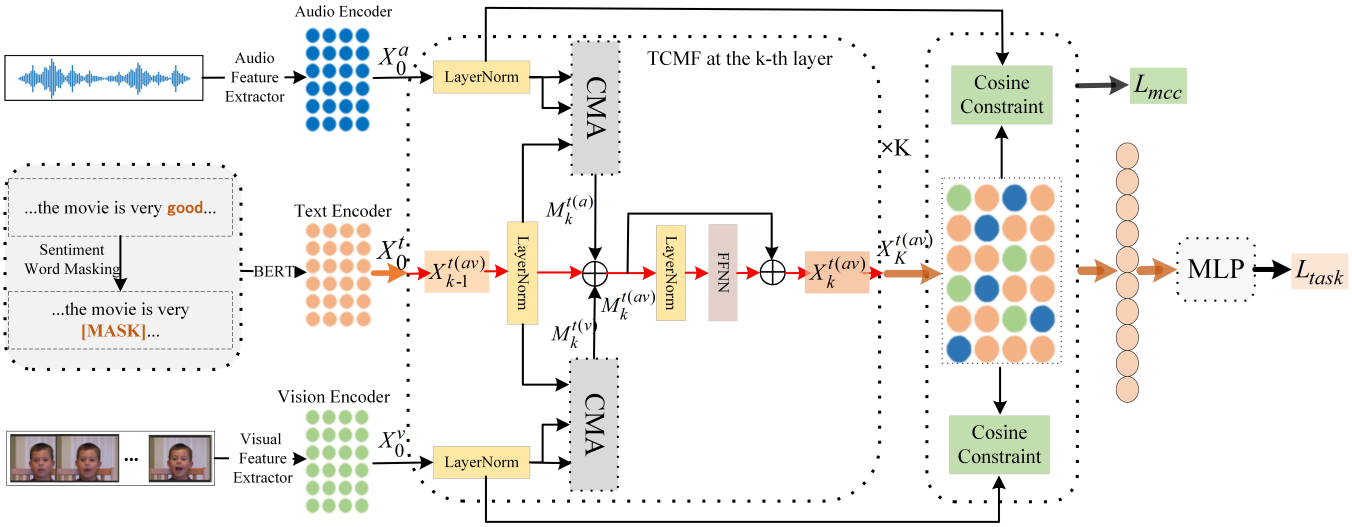

Figure 2: The overall architecture of the KEBR model. $X_0^M (M \in \{t, a, v\})$ denotes the original feature sequence before fusion. $X_k^{t(av)}$ denotes the output of TCMF at layer k. CMA denotes cross-model attention within the TCMF module. $M_k^{t(m)} (m \in \{a, v\})$ denotes the output of text doing cross-modal attention at k-th layer with audio and vision, respectively. L_{mcc} denotes multimodal cosine constraint loss. L_{task} denotes sentiment prediction loss.

fusion (LMF) [19] reduces the computational cost of TFN by utilizing a low-rank tensor. The multimodal transformer (MulT) [35] applies cross-modal attention to transform one modality to another. Different from MulT, our KEBR model injects low-level features of non-verbal information into the text feature representation through multimodal interaction. It uses the enhanced text representation as the joint representation for MSA tasks.

Given the remarkable success of pre-trained language models in NLP, MISA [12] learns invariant and specific representations of each modality. The self-supervised multitask multimodal Sentiment Analysis Network (Self-MM) [48] introduced a self-supervised label generation module for acquiring additional unimodal labels. MAG-BERT [29] introduced a multimodal adaptation gate that enables BERT to accept representations of non-verbal modalities. MMIM [11] maximizes mutual information in the multimodal fusion pipeline to preserve the fusion of task-relevant information among modalities. To explore the dynamic relationship between modalities, HyCon [22] performs comparative learning intra-modal and inter-modal to explore cross-modal interactions. Furthermore, ConKI uses the dataset from one MSA task as domain-specific knowledge trained on a different dataset, leading to the model's performance becoming overly reliant on the labeled dataset. CENet [37] utilizes visual and audio sentiment information to enrich the text representation in the pre-trained language model, while still relying on the labeled data.

Different from existing MSA models, KEBR not only uses audio and vision of MSA task labeled data to enhance the corresponding text representation but also uses sentiment knowledge from large-scale unlabeled videos to promote multimodal sentiment representation learning, which facilitates further training on limited MSA datasets.

2.3 Imbalanced Optimization

By integrating various sensory inputs, multimodal methods contribute to enhancing overall task comprehension and performance [2]. However, in practice, it has been found that even if a multimodal model outperforms its unimodal counterpart, one modality will dominate the optimization process because the unimodal converges at different speeds. Thus, causing the other modalities to be neglected and unable to play their ability [38]. To address this issue, some studies have proposed methods such as gradient mixing [38], dynamic gradient modulation [26], and an information-sharing multimodal fusion strategy [24]. MMCosine [43] proposes a multimodal cosine loss, which performs a modal L2 normalization of the features and weights. Inspired by this, we propose a multimodal cosine constrained loss function (MCC), which extends from audio-visual bimodality to address imbalance optimization among text, audio, and visual modalities compared to previous imbalance mitigation methods. Our method is self-supervised and independent of the model architecture.

3 INTRODUCTION METHODOLOGY

As shown in Fig. 2, the framework of KEBR consists of four main components: sentiment knowledge-guided masking and prediction, text-based cross-modal fusion (TCMF), multimodal cosine constraint loss function (MCC), and sentiment intensity prediction.

3.1 Task Setup

The task of MSA is to calculate the sentiment information in videos based on multimodal signals ($M = \{text(t), audio(a), visual(v)\}$). The feature sequences of these signals can be denoted as $X^t \in \mathbb{R}^{l_t \times d_t}$, $X^a \in \mathbb{R}^{l_a \times d_a}$ and $X^v \in \mathbb{R}^{l_v \times d_v}$. $l_{M \in \{t,a,v\}}$ denotes the sequence length of each modality, and d_M denotes the dimension of

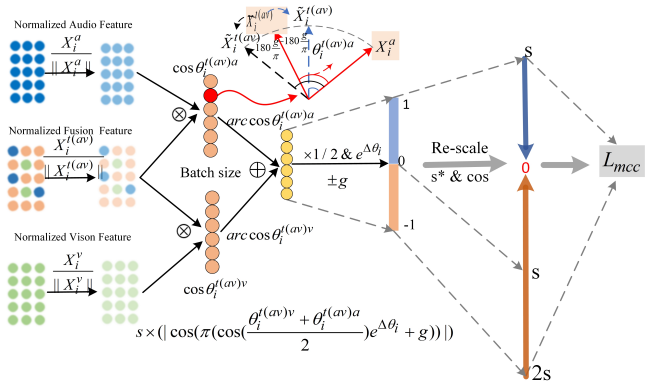

Figure 3: The process and effect of MCC. $\frac{X_i^M}{||X_i^M||}(M \in \{t,a,v\})$ **denotes normalization.** $\cos\theta_i^{t(av)m}(m \in \{a,v\})$**denotes the cosine similarity between the multimodal feature sequence** $X_K^{t(av)}$ **and** X_0^m **of sample** i **.** $\Delta\theta_i = \arccos\theta_i^{t(av)a} - \arccos\theta_i^{t(av)v}$**.** g **denotes the angle penalty used to adjust the cosine convergence value.** s **denotes the scaling factor. For samples with** $\cos\theta_i^{t(av)m} < 0$ **, it means that there is no similarity, and it is gradually optimized during training after giving it a larger loss** $s + f_{mcc}^i$ **, so we only discuss the case** $\cos\theta_i^{t(av)m} \geq 0$ **below.**

the feature. Given the multimodal sequence $X^M = \{x_1^M, x_2^M, ..., x_{l_M}^M\}$, the main tasks of KEBR are divided into two stages: pre-training and fine-tuning (testing). The pre-training task is to predict the sentiment values $y \in \mathbb{R}$ of the sentiment words in the unlabeled videos with the help of sentiment knowledge and non-verbal information. Pre-training aims to facilitate the sentiment knowledge in the unlabeled videos to promote the learning of multimodal sentiment representation. The testing task is to predict the polarity $y \in \{positive, neutral, negative\}$ or sentiment values $y \in \mathbb{R}$ of the entire labeled videos by fine-tuning the pre-trained model.

3.2 Sentiment Knowledge-guided Masking and Prediction

KEBR captures common sentiment knowledge in unlabeled videos by predicting the sentiment value of masked sentiment words, which is helpful for further learning with the limited labeled data of MSA. Specifically, given an opinion video without sentiment labels, Automatic Speech Recognition (ASR) is used to obtain the transcribed text of the video. Subsequently, the words with the highest sentiment intensity in the transcribed text are identified to be masked $y \in \{positive, neutral, negative\}$ based on a predetermined sentiment lexicon [14]. At the same time, the sentiment score of the word and its position in the text are recorded. The masked text be represented as follows.

$$T' = \{w_1, w_2, ..., w_{mask}, ..., w_n\} \tag{1}$$

where w_{mask} denotes the masked word. For text sentences that do not contain sentiment words, a word in the sentence is randomly masked, and its sentiment score is set to $y_{mask} = 0.0$.

Masked text T' with accompanying non-verbal information is used as input for KEBR pre-training. The sentiment score serves as the label to predict the sentiment value of the masked words. This prediction is assisted by text-based cross-modal fusion and a multimodal cosine constraint loss function. Thereby, word-level sentiment information is embedded into the pre-trained multimodal representation to enhance the model's performance on MSA-specific tasks.

3.3 Text-based Cross-Modal Fusion

In MSA, different non-verbal information may convey different sentiments for the same word. Therefore, the precise sentiment semantics should be determined by the word itself and the accompanying non-verbal behavior[39]. However, many studies have shown that text plays a more central role in MSA than non-verbal information [10, 12, 30, 31, 33, 37, 49] . Inspired by MulT [35], which focuses on low-level features in other modalities to fuse multimodal information, a text-based cross-modal fusion method (TCMF) is designed. In this method, low-level features from audio and visual modalities are repeatedly injected into the text feature space (K times) to enhance the multimodal representation of the text. The capacity of the fused feature parameters remains unchanged, avoiding overfitting.

In 2, the feature sequence $X_0^M(M \in \{t,a,v\})$ is used as the input to TCMF. The TCMF module consists of a total of K-th layers. The calculation for the TCMF at layer can be expressed as:

$$X_k^{t(av)} = TCMF_k(X_0^a, X_{k-1}^{t(av)}, X_0^v) \tag{2}$$

Where $0 \leq k \leq K$, if $k = 1$, $X_1^{t(av)} = TCMF_1(X_0^a, X_0^t, X_0^v)$.The main difference is that the audio and vision inputs of different layers of TCMF X_0^a are and X_0^v , while the text is input as $X_0^t(k=1)$ or $X_{k-1}^{t(av)}(1 < k \leq K)$. Non-verbal information and text within the TCMF module interact through cross-modal attention (CMA). Taking text and audio as examples, the Q (query), K (key), and V (value) of CMA are represented as follows:

$$Q^t = LN(X_{k-1}^t) \cdot W_Q^t \tag{3}$$

$$K^a = LN(X_0^a) \cdot W_K^a \tag{4}$$

$$V^a = LN(X_0^a) \cdot W_V^a \tag{5}$$

Where $W_Q^t \in \mathbb{R}^{d_t \times d_t}$, $W_K^a \in \mathbb{R}^{d_a \times d_t}$ and $W_V^a \in \mathbb{R}^{d_a \times d_t}$ are weighs. $X_{k-1}^t = X_0^t(k=1)$, $X_{k-1}^t = X_{k-1}^{t(av)}(1 < k \leq K)$. Further, the CMA for audio and text can be expressed as follows.

$$\begin{aligned} M_k^{t(a)} &= CMA(Q^t, K^a, V^a) \\ &= \text{softmax}(\frac{Q^t \cdot K^a}{\sqrt{d_t}}) \cdot V^a \end{aligned} \tag{6}$$

Where $M_k^{t(a)} \in \mathbb{R}^{N \times d_t}$ denotes the text sequence enhanced by audio modal injection. Likewise, the text sequence enhanced by audio modal injection can be denoted as $M_k^{t(v)}$. $M_k^{t(a)}$ and $M_k^{t(v)}$ are combined with the output $X_{k-1}^{t(av)}$ of TCMF at the layer $k-1$ to obtain a text representation enhanced by non-verbal information.

$$M_k^{t(av)} = M_k^{t(a)} + LN(X_{k-1}^{t(av)}) + M_k^{t(v)} \tag{7}$$

Where $M_k^{t(av)} \in \mathbb{R}^{N \times d_t}$ denotes a text representation that injected audio and vision. Specifically, since $Q^t\,(k>1)$ is the result of the TCMF output from the previous $k-1$ layer, also contains the injected visual information of the former layers. However, to differentiate $M_k^{t(a)}$ from $M_k^{t(av)}$, only $[M_k^{t(a)}$ is used to signify the fusion of audio and text at the $k-th$ layer.

$M_k^{t(av)}$ is processed by layer normalization, feedforward neural network (FFNN), and residual connection. The $k-th$ layer of non-verbal information-enhanced text representation $X_k^{t(av)}$ is obtained.

$$X_k^{t(av)} = FFNN(LN(M_k^{t(av)})) + M_k^{t(av)} \tag{8}$$

Where $X_k^{t(av)} \in \mathbb{R}^{N \times d_t}$ denotes the output of the TCMF module at layer k.

3.4 Multimodal Cosine Constraint

In Section 3.3, we employ the low-level features of both audio and visual modalities to enhance the multimodal representation of the text via TCMF. However, in the process of information fusion, we find that when a non-verbal modality gains a relative advantage in the fusion, it tends to quickly strengthen its advantages, thus precluding the fusion of other non-verbal information within the main modality. To optimize the imbalance of modality in the multimodal joint representation, we design a multimodal cosine constraint loss function (MCC), which constrains the fusion of non-verbal information by utilizing the cosine constraint. This approach facilitates the synchronous convergence of different non-verbal information and exploits the capabilities of different modalities.

As shown in Fig. 3, the sample feature sequences are firstly normalized by L2. Subsequently, the cosines of the fused multimodal features $X_K^{t(av)}$ and the original features X_0^a , and X_0^v are calculated, respectively.

$$\cos\theta_i^{t(av)a} = \frac{X_{K(i)}^{t(av)^T} X_{0(i)}^a}{\| X_{K(i)}^{t(av)} \| \cdot \| X_{0(i)}^a \|} \tag{9}$$

where $\cos\theta_i^{t(av)a}$ represents the cosine of $X_K^{t(av)}$ and X_0^a for the sample i . Similarly, the cosine $X_K^{t(av)}$ and X_0^v can be expressed as $\cos\theta_i^{t(av)v}$. Given n samples, the MCC loss function can be expressed as follows.

$$L_{mcc} = \frac{1}{n}\sum_{i=0}^{n} f_{mcc}^i \tag{10}$$

$$f_{mcc}^i = s \times (|\cos(\pi(\cos(\frac{\theta_i^{t(av)v} + \theta_i^{t(av)a}}{2})e^{\Delta\theta_i} + g))|) \tag{11}$$

Where $\Delta\theta_i = \arccos\theta_i^{t(av)a} - \arccos\theta_i^{t(av)v}$, f_{sim}^i denotes the MCC loss function of the sample i . g is the angle penalty used to adjust the cosine of $X_K^{t(av)}$ with X_0^a and X_0^a when the similarity loss converges. Due to the function $|\cos(x)| \in [0,1]$, the introduction of parameters to balance the size loss of MCC and the main task loss ensures the convergence of the model.

$$y(x_i) = \cos(\frac{\theta_i^{t(av)v} + \theta_i^{t(av)a}}{2})e^{(\arccos\theta_i^{t(av)a} - \arccos\theta_i^{t(av)v})} \tag{12}$$

When f_{mcc}^i converges, the value of the cosine function $y(x_i)$ value is assumed to be $\hat{y}(x_i)$. To induce $y(x_i)$ convergence $\hat{y}(x_i)$, both $\theta_i^{t(av)a} \to \arccos(\hat{y}(x_i))$ and $\theta_i^{t(av)v} \to \arccos(\hat{y}(x_i))$ are required. Meanwhile, $\arccos\theta_i^{t(av)a} - \arccos_i^{t(av)v}$ must be minimized, i.e., $\arccos\theta_i^{t(av)a}$ and $\arccos\theta_i^{t(av)v}$ must be very close to each other. Under this cooperative constraint, the convergence of non-verbal information should be synchronous to avoid modal bias.

The non-verbal modalities are expected to play an auxiliary and augmenting role in the multimodal fusion. To maintain the dominance of the text master modality, it should prevent the cosine function $\cos\theta_i^{t(av)m}(m \in \{a,v\}) \to 0$ or $\cos\theta_i^{t(av)m}(m \in \{a,v\}) \to 1$, i.e., the cosine values of the fused feature sequences and the non-verbal modal feature sequences being very high or very low. To achieve this, we reintroduce the cosine function, resulting a higher loss for the similarity score $\cos\theta_i^{t(av)m}(m \in \{a,v\}) \to 0$ or $\cos\theta_i^{t(av)m}(m \in \{a,v\}) \to 1$, as shown in table 3. Simultaneously, an angle penalty g is introduced to regulate the convergence value of the cosine function $y(x_i)$. Through this joint regulation, the convergence of the multimodal cosine constrained loss f_{mcc}^i for the sample i can be expressed as follows.

$$f_{mcc}^i = \lim_{(y(x_i)+g)\to\hat{y}(x_i)} s \times (|\cos(\pi(y(x_i)+g))|) = 0 \tag{13}$$

3.5 Sentiment Intensity Prediction

In the pre-training phase of KEBR, the main task is to predict the intensity of sentiment words that have been masked. This is to use the sentiment knowledge from large, unlabeled videos to improve multimodal sentiment representation learning. Masked text encoded by BERT [9] and enhanced with low-level features from audio and visual modalities by TCMF, can be represented as follows:

$$X_K^{t(av)} = \{x_{K,1}^{t(av)}, x_{K,2}^{t(av)}, ..., x_{MASK}^{t(av)}, ..., x_{K,n}^{t(av)}\} \tag{14}$$

Where $x_{MASK}^{t(av)}$ denotes the masked word that has been reinforced by non-verbal information. To predict the sentiment value of the masked word, we employ a multilayer perceptron (MLP) with a ReLU activation function acting as a classifier.

$$y_{pred} = MLP_{\theta_{FC}}(x_{MASK}^{t(av)}) \tag{15}$$

Where θ_{FC} denotes the parameters of the fully connected network, and y_{pred} is the predicted sentiment value.

$$L_{task} = \frac{1}{n}\sum_{i=1}^{n} |y_{pred} - y_{MASK}| \tag{16}$$

$$L = L_{task} + L_{mcc} \tag{17}$$

where n is the batch size. L is the total loss of the model pre-training. L_{task} is the masked sentiment word sentiment value prediction loss.

Our task in the testing phase is also affective computing. Therefore, in the testing phase, we only added an output layer to the pre-trained language model as well as the multimodal fusion module to generate task-specific predictions. Then, we fine-tuned the labeled multimodal dataset to validate the performance of KEBR in the MSA task.

Table 1: Results on MOSI and MOSEI. In Acc-2 and F1, the left of the "/" is "negative/nonnegative" and the right is "negative/positive". "*" denotes the result is from ConKI [49]. "Δ" denotes the reimplementation of the non-verbal feature mentioned in 4.2 .Mark the best results in bold.

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LFN*	-/80.8	-/80.7	34.9	0.901	0.698	-/82.5	-/82.1	50.2	0.593	0.7
LMF*	-/82.5	-/82.4	33.2	0.917	0.695	-/82.0	-/82.1	48	0.677	0.623
MISA*	80.79/82.10	80.77/82.03	-	0.804	0.764	82.59/84.23	82.67/83.97	-	0.548	0.724
MAG-BERT*	82.37/84.43	82.50/84.61	43.62	0.781	0.727	82.51/84.82	82.77/84.71	52.67	0.543	0.755
HyCon*	-/85.2	-/85.1	46.6	0.713	0.79	-/85.4	-/85.6	52.8	0.601	0.776
MIMM*	84.14/86.06	84.00/85.98	46.65	0.7	0.8	82.24/85.97	82.66/85.94	54.24	0.526	0.772
ConKI*	84.37/86.13	84.33/86.13	**48.43**	**0.681**	0.816	82.73/86.25	83.08/86.15	54.25	0.529	0.782
MulTD$^{\Delta}$	79.51/80.47	79.46/80.49	36.74	0.892	0.667	81.10/83.63	81.05/83.46	52.34	0.605	0.671
CENETD$^{\Delta}$	83.57/85.31	83.54/85.29	44.86	0.725	0.799	83.49/86.33	83.74/86.21	54.15	0.529	0.773
Self-MMD$^{\Delta}$	83.21/85.21	83.18/85.18	44.65	0.701	0.774	82.93/84.07	83.14/84.12	53.74	0.554	0.75
KEBR	**84.84/87.27**	**84.83/87.25**	47.81	0.683	**0.819**	**84.01/86.74**	**84.25/86.68**	**54.37**	**0.517**	**0.799**

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Pre-training datasets. KEBR was pre-trained on the VoxCeleb[1] dataset, which is a sizable dataset of English speaker recognition with rich sentiment. The dataset has two parts: VoxCeleb1 [23] and VoxCeleb2 [7]. VoxCeleb1 has over 100,000 discourses from 1251 celebrities from YouTube videos. VoxCeleb2 has more than one million discourses from 6112 speakers. According to previous research [41], video clips without English were excluded.

Fine-tuning datasets. We fine-tuned the pre-trained KEBR on two benchmark datasets in the MSA field: MOSI [52] and MOSEI [1]. The details of the datasets are shown in Appendix A.

Following previous works [7, 12, 22, 33, 49], we present our experimental results in both regression and classification. For regression, we present the mean absolute error (MAE) and Pearson correlation (Corr). For classification, we calculate the Acc-2 and F1 scores for both the negative positive (zero excluded) and non-negative positive (zero included) settings. Additionally, In addition, we calculate Acc-7, which shows the percentage of predictions correctly classified into seven intervals between -3 and +3. Except for MAE, higher values indicate better performance for all metrics.

4.2 Experimental Design

For the text, audio, and visual information in the dataset, we use BERT [9], librosa [4], and OpenFace [3] to perform the feature extraction of the corresponding information. To be consistent with the sentiment values of the test dataset, we linearly scale the sentiment scores shown in the sentiment lexicon [14] from [-4, +4] to [-3, +3].

Training Details: All models are built on the Pytorch [25] toolbox with the NVIDIA RTX A100 GPU. The batch size is set to 32 and the epoch is set to 200. The initial learning rate is set to 5e-6 for BERT and 1e-4 for other parameters. Adam[16] is used as the optimizer. The angle penalty g is 0.2. Appendix B describes the details of video feature extraction and hyper-parameter setting.

4.3 Baselines

Given that the pre-training dataset lacks explicit word timestamps, our model does not need to align text words with vision and audio. We performed a comprehensive comparative analysis of KEBR with baselines and state-of-the-art models on an unaligned setup, with the comparative model as follows: TFN [50], LMF [19], MISA [12], MAG-BERT [29], HyCon [22], MMIM [11], ConKI [49], MulT [35], CENet [37], Self-MM [48]. Please refer to Appendix C for more details on these baselines.

4.4 Quantitative Results and Analysis

Following previous works [8, 12, 22, 49], we run our model five times with the same hyper-parameter settings and report the average performance of all the metrics in Table 1. It is evident that KEBR produces better or more competitive results in MOSI and MOSEI.

In the experiments based on the re-implementation of non-verbal feature extraction in Section 4.2 (Marked with "Δ" in Table 1), our KEBR significantly outperforms the other models on all experimental metrics for two datasets. The baseline code is from publicly available repositories[2].

Compared with the experimental metrics in the latest research papers, KEBR achieved better or competitive results, except for Acc-7 and MAE on MOSI are slightly lower than ConKI. This may be because, ConKI utilizes adapters to inject a larger amount of data from MOSEI as MOSI-specific domain knowledge into the learning process. However, KEBR was pre-trained on the unlabeled dataset VoxCeleb to learn more general sentiment patterns, so the metrics of the more refined multi-classification tasks were slightly lower than ConKI. However, on the MOSEI dataset, KEBR outperforms ConKI in all metrics, with an average performance improvement of 1.72%. Considering that the size of MOSEI is much larger than

[1]https://mm.kaist.ac.kr/datasets/voxceleb

[2]https://github.com/thuiar/MMSA

Table 2: Results on MOSE and MOSEI with different amounts of pre-training data. "Avg" donates the average improvement across all metrics. "↑"means increase, and "↓" means decrease.

Dataset	Pre-train	Acc-2	F1	Acc-7	MAE	Corr	Ave
MOSI	VoxCeleb1	84.4/87.2	84.32/87.19	47.1	0.702	0.808	**0.99%**↑
	VoxCeleb2	84.84/87.27	84.83/87.25	47.81	0.683	0.819	
MOSEI	VoxCeleb1	83.67/86.21	83.99/86.16	53.33	0.534	0.776	**1.37%**↑
	VoxCeleb2	84.01/86.74	84.25/86.68	54.37	0.517	0.799	

Table 3: Results on MOSE and MOSEI with different pre-trained language models. "Avg" donates the average improvement across all metrics. "↑"means increase, and "↓" means decrease.

Dataset	Bert	Acc-2	F1	Acc-7	MAE	Corr	Avg
MOSI	Bert-base	83.84/86.14	83.82/86.18	45.59	0.734	0.795	**1.88%**↑
	Bert-large	84.4/87.2	84.32/87.19	47.1	0.702	0.808	
MOSEI	Bert-base	83.6/85.1	83.76/85.28	51.88	0.554	0.751	**1.79%**↑
	Bert-large	83.67/86.21	83.99/86.16	53.33	0.534	0.776	

Table 4: Ablation experiments on MOSI. Pre-t means pre-training. TCMF means text-based cross-modal fusion. MCC means multimodal cosine constraint. "✓" means with and "×" means without. Excluding TCMF, only removing the text-based part and retaining the overall network architecture of cross-modal fusion. The Avg of C2 C5 is calculated relative to C1 as the benchmark. "Avg" represents the average improvement across all metrics. "↑"means increase, and "↓" means decrease.

No.	Pre-t	TCMF	MCC	Acc-2	F1	Acc-7	MAE	Corr	Avg
C1	✓	✓	✓	84.4/87.2	84.32/87.19	47.1	0.702	0.808	-
C2	×	✓	✓	84.11/86.74	84.03/86.73	46.76	0.705	0.803	**0.51%** ↓
C3	✓	×	✓	81.92/83.99	81.82/83.96	38.48	0.826	0.737	**8.3%** ↓
C4	✓	✓	×	83.75/86.28	83.69/83.29	45.7	0.7069	0.8074	**1.55%** ↓
C5	×	×	×	80.9/82.32	80.88/82.35	34.4	0.927	0.7302	**12.58%** ↓

that of MOSI, injecting knowledge from MOSEI into MOSI has a greater impact on the downstream task than injecting knowledge from MOSI into MOSEI [49]. This is consistent with the research results of this paper that utilizing external data as an injection of sentiment knowledge facilitates model training. The difference is that our model KEBR utilizes unlabeled data from the network whereas ConKI uses domain-specific (MSA) labeled data.

To study the impact of learning from external data on the model, we pre-trained KEBR using the VoxCeleb1 (132K) and VoxCeleb2 (947K) datasets.

The results are presented in Table 2. A substantial amount of pre-training data results in a more significant performance improvement. Compared with VoxCeleb1, the average performance of MOSI and MOSEI pre-trained on VoxCeleb2 is improved by 0.99% and 1.37%, respectively. The performance improvement of MOSEI exceeds that of MOSI, possibly due to the relatively small size of the MOSI dataset, which might have been more effectively enhanced using the VoxCeleb1 pre-trained model. It may also contain labels with noise, which has been found in some previous studies [37]. In Section 4.6, when the effects of different g on performance are analyzed on two data sets, it is further demonstrated that the model trained on MOSI has greater instability than MOSEI.

To investigate the impact of backbone language models with different parameters on KEBR, we conducted comparative analyses on VoxCeleb1 using Bert-base and Bert-large models with 110M and 340M parameters in MOSI and MOSEI.

As shown in Table 3, Bert-large enhances the average performance of MOSI and MOSEI by 1.88% and 1.79%, respectively, compared to Bert-base. By comparing Tables 2 and 3, we observe that larger language models may outperform models trained on larger pre-training datasets. Bert-large has stronger language comprehension and representation ability. This is consistent with the earlier conclusion [5] that scaling up can significantly enhance performance. Moreover, the text encoded by a larger pre-trained language model can acquire a broader representation space for non-verbal information under the same cosine constraint, while retaining the text features, thereby facilitating further improvement in performance.

4.5 Ablation Study

We conduct an associated ablation study in the MOSI dataset. Pre-training in VoxCeleb2 is costly, so we conduct an ablation analysis in the VoxCeleb1 dataset. Bert-large serves as the backbone model.

In this paper, the performance of the KEBR is primarily influenced by a sentiment knowledge-enhanced pre-training task (Pre-t) and two model components: text-based cross-modal fusion (TCMF)

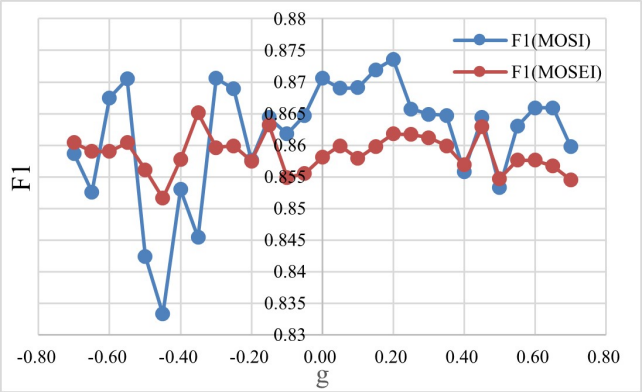

Figure 4: Analyze the angle penalty g. The blue and red curves depict the change in F1 score with varying angle penalty values on the MOSI and MOSEI datasets.

and multimodal cosine constraints (MCC). C1 represents the result of pre-training on VoxCeleb1 followed by fine-tuning based on the pre-training model. C2 represents no pre-training, with tests conducted directly on the initial BERT and TCMF parameters. C3 represents the pre-training and testing phases that do not involve text-based multimodal fusion. C4 means that no multimodal cosine constraints are applied on both pre-training and test tasks.

We evaluated the impact of not including the KEBR components on performance using the average degradation rate of all the metrics. As shown in Table 4, it can be observed that excluding different parts of the KEBR will lead to performance degradation. Further analysis shows that without TCMF(C2) KEBR performance decreases by 8.3%. This decrease is significantly larger compared to the exclusion of Pre-t(C2) and MCC(C4). This suggests that a text-based approach in the MAS task can significantly improve model performance. The average performance of KERB without Pre-t(C2) only decreases by 0.51%, which may be attributed to conducting our ablation experiments on the smaller VoxCeleb1 dataset (132 K). Comparing the results in Table 2 to VoxCeleb2 (947 K), the overall performance without Pre-t(C2) drops to 1.49%. This suggests that the impact of pre-training on model performance is limited by the amount of pre-training data. Meanwhile, we observed that the average performance without MCC(C4) decreased by 1.55%, even surpassing the 1.49% decline observed with the larger pre-training dataset VoxCeleb2. This suggests the presence of a potential modal imbalance issue in MSA, and incorporating additional constraints of MCC to balance the modal representation can significantly enhance the model's performance.

4.6 Further Analysis

We analyze the effects of different g on the model and the imbalance phenomena in the MSA task (provided by Appendix D.1), and further, visualize the reasons for the imbalance phenomena in multimodal fusion(provided by Appendix D.2).

Analyze the angle penalty g. In Section 4.5, experiments demonstrate that MCC is effective and necessary in KEBR. Next, we analyze the effect of different values of the angle penalty g in MCC on the MOSI and MOSEI datasets. The pre-training dataset used is VoxCeleb1.

It is observed in Fig. 4 that the variation trend of F1 on the two datasets with g is similar. However, MOSI exhibits relatively higher instability, likely due to its smaller dataset size and the potential presence of noisy labels. For $g \in [-0.5, -0.05]$, F1 is low and the performance is very unstable. For $g \in [-0.05, 0.25]$, F1 is relatively high and stable. For $g \in [0.25, 0.5]$, F1 is relatively low and unstable. Further analysis with Eq.11, d considering the periodic nature of the cosine function, suggests that the variation of F1 with g should also display a weak periodicity with a period of 1. Due to the nonlinearity of the cosine function, g varies within the same step interval, resulting in differing effects on the cosine values. For $g \in [-0.5, 0]$, the same change has a greater impact on F1 than that $g \in [0, 0.5]$, so there is a tendency for F1 to be destabilized. Furthermore, when $g \in [-0.5, 0]$, the cosine function needs to converge to a larger cosine value. This will make the proportion of non-verbal info in the fused features larger, which destroys the dominance of the text modality and causes performance degradation and instability. On the contrary, when $g \in [0, 0.5]$, the value that $y(x_i)$ needs to converge decreases as g increases. The fused feature has less non-verbal information and cannot utilize the ability of all modalities, so the performance will be decreased. However, in this scenario, text still dominates, resulting in a slight decline in performance. Considering different non-verbal information may convey different sentiments for the same word, non-verbal dominant modal fusion tends to be more unstable than text. Thus, we can guess boldly that for different datasets, when $g \in [0, 0.25]$, it will bring a large performance improvement while maintaining a robust multimodal joint representation.

To better comprehend the modal imbalance issue in MSA, a comparative experiment is conducted in Appendix D.1 to observe the convergence of various modalities at different learning rates. In Appendix D.2, modal feature distribution is visualized and analyzed with and without MCC to elucidate the underlying cause of the modal imbalance problem.

5 CONCLUSION

In this paper, we propose KEBR, which leverages the sentiment knowledge from large-scale unlabeled videos to enhance multimodal sentiment representation learning. This approach facilitates learning on limited MSA datasets. KEBR highlights the dominance of text modality in MSA tasks by employing multi-layer cross-modal cross-fusion to inject non-verbal information into text representations to enhance multimodal representation in text. Furthermore, we design the multimodal cosine constrained loss to optimize the imbalance of unimodal in joint representation and utilize the capability of different modalities. We validate the effectiveness of the KEBR model and the importance of each module through comprehensive experiments and ablation studies conducted on two benchmark datasets. The experiments show that the larger the amount of unlabeled video data and the stronger the language modeling ability, the better the performance. We believe this work can inspire further exploration of multimodal representations for various scenarios and tasks, as well as investigations into optimizing modal imbalance in joint learning.

REFERENCES

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[4] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra (Eds.). 18 – 24. https://doi.org/10.25080/Majora-7b98e3ed-003

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.723

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014). https://doi.org/10.48550/arXiv.1406.1078

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*. 1086–1090. https://doi.org/10.21437/Interspeech.2018-1929

[8] Wenliang Dai, Samuel Cahyawijaya, Yejin Bang, and Pascale Fung. 2021. Weakly-supervised multi-task learning for multimodal affect recognition. *arXiv preprint arXiv:2104.11560* (2021). https://doi.org/10.48550/arXiv.2104.11560

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). https://doi.org/10.48550/arXiv.1810.04805

[10] Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3394–3402. https://doi.org/10.1145/3503161.3548137

[11] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 9180–9192. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.723

[12] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131. https://doi.org/10.1145/3394171.3413678

[13] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4604–4614. https://doi.org/10.18653/v1/2020.emnlp-main.372

[14] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[15] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6975–6988. https://doi.org/10.18653/v1/2020.emnlp-main.567

[16] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). https://doi.org/10.48550/arXiv.1412.6980

[17] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019). https://doi.org/10.48550/arXiv.1907.11692

[19] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2247–2256. https://doi.org/10.18653/v1/P18-1209

[20] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 481–492. https://doi.org/10.18653/v1/P19-1046

[21] Sijie Mai, Songlong Xing, and Haifeng Hu. 2019. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia* 22, 1 (2019), 122–137. https://doi.org/10.1109/TMM.2019.2925966

[22] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* (2022). https://doi.org/10.1109/TAFFC.2022.3172360

[23] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*. 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

[24] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems* 34 (2021), 14200–14213.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[26] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8238–8247. https://doi.org/doi.org/10.48550/arXiv.2203.15332

[27] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 43–54. https://doi.org/10.18653/v1/D19-1005

[28] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2539–2544. https://doi.org/10.18653/V1/D15-1303

[29] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359. https://doi.org/10.18653/v1/2020.acl-main.214

[30] Piao Shi, Min Hu, Fuji Ren, Xuefeng Shi, and Liangfeng Xu. 2022. Learning modality-fused representation based on transformer for emotion analysis. *Journal of Electronic Imaging* 31, 6 (2022), 063032–063032. https://doi.org/10.1117/1.JEI.31.6.063032

[31] Piao Shi, Min Hu, Xuefeng Shi, and Fuji Ren. 2024. Deep Modular Co-Attention Shifting Network for Multimodal Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 4 (2024), 1–23. https://doi.org/10.1145/3634706

[32] Yusheng Su, Xu Han, Zhengyan Zhang, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. Cokebert: Contextual knowledge selection and embedding towards enhanced pre-trained language models. *AI Open* 2 (2021), 127–134. https://doi.org/10.1016/j.aiopen.2021.06.004

[33] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. 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. 8992–8999. https://doi.org/10.1609/aaai.v34i05.6431

[34] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4067–4076. https://doi.org/10.18653/V1/2020.ACL-MAIN.374

[35] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558. https://doi.org/10.18653/v1/p19-1656

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[37] Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia* (2022). https://doi.org/10.1109/TMM.2022.3183830

[38] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12695–12705. https://doi.org/10.48550/arXiv.1905.12681

[39] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223. https://doi.org/10.1609/aaai.v33i01.33017216

[40] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. 2022. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579* (2022). https://doi.org/10.48550/arXiv.2208.09579

[41] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2019. Utterance-level Aggregation for Speaker Recognition in the Wild. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5791–5795. https://doi.org/10.1109/ICASSP.2019.8683120

[42] Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). https://doi.org/10.1109/TPAMI.2023.3275156

[43] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. 2023. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096655

[44] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. 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)*. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421

[45] Kaicheng Yang, Hua Xu, and Kai Gao. 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 521–528. https://doi.org/10.1145/3394171.3413690

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[47] Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 3695–3706. https://doi.org/10.18653/V1/2020.ACL-MAIN.341

[48] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. 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. 10790–10797. https://doi.org/10.1609/aaai.v35i12.17289

[49] Yakun Yu, Mingjun Zhao, Shi ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis. Findings of the Association for Computational Linguistics: ACL 2023 (2023), 13610–13624. https://doi.org/10.48550/arXiv.2306.15796

[50] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115

[51] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. https://doi.org/10.1609/aaai.v32i1.12021

[52] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. abs/1606.06259 (2016). https://doi.org/10.48550/arXiv.1606.06259

[53] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1441–1451. https://doi.org/10.18653/v1/P19-1139

[54] Qinghua Zhao, Shuai Ma, and Shuo Ren. 2022. KESA: a knowledge enhanced approach for sentiment analysis. *arXiv preprint arXiv:2202.12093* (2022). https://doi.org/10.48550/arXiv.2202.12093

[55] Yang Zhao, Tetsuya Nasukawa, Masayasu Muraoka, and Bishwaranjan Bhattacharjee. 2023. A Simple Yet Strong Domain-Agnostic De-bias Method for Zero-Shot Sentiment Classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3923–3931. https://doi.org/10.18653/v1/2023.findings-acl.242

[56] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5218–5230. https://doi.org/10.18653/V1/2020.ACL-MAIN.466