Contents lists available at ScienceDirect

Expert Systems With Applications

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

A fine-grained modal label-based multi-stage network for multimodal sentiment analysis

Junjie Peng [a,b,*], Ting Wu [a], Wenqiang Zhang [c,d,**], Feng Cheng [e], Shuhua Tan [f,g], Fen Yi [f,g], Yansong Huang [a]

[a] *School of Computer Engineering and Science, Shanghai University, Shanghai, China*
[b] *Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China*
[c] *Academy for Engineering and Technology, Fudan University, Shanghai, China*
[d] *School of Computer Science and Technology, Fudan University, Shanghai, China*
[e] *Hasso Plattner Institute, University of Potsdam, Potsdam, Germany*
[f] *National Engineering Laboratory for Logistics Information Technology, Shanghai, China*
[g] *YTO Express Co., Ltd., Shanghai, China*

ARTICLE INFO

Keywords:
Sentiment analysis
Multimodality
Multi-task learning
Multimodal fusion

ABSTRACT

Sentiment analysis is a challenging but valuable research topic in affective computing. It can improve the quality of various real-world applications, including financial market prediction, disease analysis even politics. As sentiment may be expressed by text, image, audio, video, etc., multimodal sentiment analysis has emerged to capture information in multiple ways. Take video as an example, the analysis process may be difficult since the modalities in the video are heterogeneous and may express different sentiments. To deal with such issues, a Fine-grained modal label-based Multi-Stage Network (FmlMSN) is proposed. Utilizing seven sentiment labels in unimodal, bimodal and trimodal, the model focuses on information at different granularities from text, audio, image and the combinations of them. Meanwhile, inspired by the idea of stacking ensemble learning which is still limited in sentiment analysis, multi-stage training is performed for base learners of acoustic-visual, visual-textual and acoustic-textual. In each stage, the singleton modality and pair-wise modalities are interconnected by hard parameter sharing multi-task learning. Subsequently, the hidden bimodal features are used to train the meta-learner for the final sentiment prediction. Extensive experiments on three public datasets, including one in Chinese and two in English indicate that our model outperforms the existing state-of-the-art methods. Furthermore, empirical analysis suggests that the model is flexible and can reduce training time and calculation to some extent.

1. Introduction

Sentiment analysis is important as it can provide useful feedback on human emotions. It is a basic task to achieve intelligent human–computer interaction (Yadollahi et al., 2017). Since traditional textual sentiment analysis is ineffective in social media applications, multimodal analysis has become a hot research topic. In this study, each source or form of information is regarded as a modality and multiple modalities are used together to solve problems. For example, videos contain a large amount of data in form of text, audio, and video. Text (T) carries the semantic information of the spoken sentence, audio (A) reveals speaker's speed, intonation, and emphasis on some words, video (V) reflects the gesture and posture of the speaker. Making full utilization of these modalities can better understand human sentiments and recognize human intentions.

Generally, there are two main challenges in processing multimodal data. One is intra-modal temporal information extraction, and the other is cross-modal interaction (Yu, Jiang et al., 2020). As a relatively simple problem, intra-modal information extraction has been widely studied and many promising achievements have been obtained such as LSTM (Hochreiter & Schmidhuber, 1997), CNN (Karpathy et al., 2014) and attention mechanism (Vaswani et al., 2017) and so on. However, cross-modal interaction is still an urgent problem that needs to solve.

Compared with intra-modal information extraction, cross-modal interaction is far more difficult, as the heterogeneity of various modalities

[*] Corresponding author at: School of Computer Engineering and Science, Shanghai University, Shanghai, China.
[**] Corresponding author.

E-mail addresses: jjie.peng@shu.edu.cn (J. Peng), wuting1972@shu.edu.cn (T. Wu), wqzhang@fudan.edu.cn (W. Zhang), feng.cheng@hpi.uni-potsdam.de (F. Cheng), 00000226@yto.net.cn (S. Tan), 00001225@yto.net.cn (F. Yi), rockpine@shu.edu.cn (Y. Huang).

https://doi.org/10.1016/j.eswa.2023.119721
Received 18 November 2021; Received in revised form 10 February 2023; Accepted 17 February 2023
Available online 21 February 2023
0957-4174/

increases the difficulty of modal fusion (Mai et al., 2021). Traditional ways to solve this problem are to align multiple modalities before fusion, like forcing each modality to align at the word level through manual preprocessing (Poria, Peng et al., 2017; Zadeh, Liang, Mazumder et al., 2018; Zadeh, Liang, Poria et al., 2018). In this way, visual and acoustic features are averaged across the time interval to align with each word. Different from the word-level alignment, Tsai et al. (2019) proposed a cross-modal attention mechanism to achieve semantic alignment and capture the long-term dependencies of features.

The above modality-aligned methods are evaluated to be effective for video sentiment analysis on MOSEI (Zadeh, Liang, Poria et al., 2018), CMU-MOSI (Zadeh et al., 2016), and IEMOCAP (Busso et al., 2008) datasets. However, they are limited to simple scenarios because most of the video clips in these datasets are monologue videos or dialogue videos shot indoor. In other words, the environments are easy to manage with specific figures and contextual relationships that the end-to-end models can achieve better performance. However, in movie and TV series, the shooting environments are more real-life where the unalignment of modal data is very common. Besides, the sentiment orientation in each modality is different so that makes the importance of each modality variable. The meaning of words and sentences spoken by speakers often changes dynamically according to the non-verbal behaviors (Cao et al., 2021). For example, if someone said ‘It is OK’, it could express negative sentiment if it is accompanied by a forced smile or suppressed voice. In the circumstances, designing a robust model to mine the interaction relations between various modalities is more challenging than that in simple scenarios. Researchers tend to solve the problems implicitly by utilizing attention mechanism in an end-to-end manner. One reason is that CMU-MOSI and MOSEI only contain unified multimodal labels (i.e., the general sentiment of the video), which cannot always represent the independent sentiment of single modality or paired modalities. Recently, unimodal labels and multimodal labels for CH-SIMS have been provided, encouraging research to explore the differences between modalities further. However, the research is suppressed by exploring bimodal sentiment as bimodal labels are unavailable in almost all multimodal datasets.

To detect the contribution of different levels of sentiment, we annotate bimodal labels on CH-SIMS and unimodal labels and bimodal labels on CMU-MOSI. Meanwhile, we propose a Fine-grained modal label-based Multi-Stage Network (FmlMSN), which takes the sentiment differences between different modalities into account and introduces unimodal, bimodal, and multimodal sentiment labels. Additionally, inspired by the stacking ensemble learning which typically considers heterogeneity and diversity (Kazmaier & van Vuuren, 2022), we train the learners of acoustic-visual (AV), visual-textual (VT) and acoustic-textual (AT) stage by stage. Moreover, to explore the relations between modalities, Multi-Task Learning (MTL) is introduced in each stage as it can improve the generalization of a model like people often learn a new task by applying the knowledge from previous tasks (Zhang & Yang, 2017). It introduces joint learning to achieve better performance. That is, one task is used as the main task and other tasks as auxiliary tasks (Sener & Koltun, 2018). Following the concept of joint learning, many researchers have focused on sentiment analysis in multimodal and a lot of promising results have been achieved (Akhtar et al., 2019; Fortin & Chaib-draa, 2019; Hazarika et al., 2020; Tian et al., 2018; Wu et al., 2020; Yu, Xu et al., 2020). These studies consider different tasks including sentiment intensity prediction, emotion recognition, etc. Some of them can not only analyze the single modality, but also learn multiple modalities jointly. In this way, they can learn more general feature representations by ignoring the noise carried by different tasks. In this paper, the bimodal sentiment analysis acts as the main task while the unimodal sentiment analysis as the auxiliary task. They are interconnected by feature sharing and loss constraint. After that, the obtained bimodal features are concatenated as the input of the meta-learner for the final sentiment prediction.

The main contributions can be summarized as follows:

- Bimodal sentiment labels are annotated on two datasets to explore the contribution of different hierarchical modalities sentiment, as bimodal labels in multimodal dataset are not considered and unavailable.
- A fine-grained modal label-based Multi-Stage Network for sentiment analysis is proposed which means the method treats modalities of different granularities as independent tasks for multi-stage training.
- As a more fine-grained sentiment analysis method, the method considers the sentiment differences of unimodal, bimodal, and multimodal which makes it be more effective in sentiment analysis than other models do.

The remainder of this paper is organized as follows. Section 2 reviews the literature of multimodal sentiment analysis. Section 3 describes the extension of the multimodal sentiment datasets. Section 4 describes the proposed model in detail. Section 5 reports the settings and results of the experiments. Section 6 concludes the paper.

2. Related work

As an important research topic in affective computing, Natural Language Processing (NLP), sentiment analysis has been widely studied. In the early works, sentiment analysis mainly focuses on text-based study and this kind of study generally includes two approaches: lexicon-based and machine learning-based (Kaur & Kautish, 2019). In lexicon-based sentiment analysis, both sentiment dictionary and opinion words are used to match them with data to pronounce the polarity. While in machine learning-based sentiment analysis, supervised learning approaches such as decision tree classifier, linear classifier, and rule-based classifier are used to categorize text. With the development of speech-based emotion recognition, researchers have proposed methods combining both text and audio which have achieved better performance than unimodal system (Sahu et al., 2019).

Since facial expressions are proved to be vital for emotion detection, it has been widely used as a modality to provide more clues (Chen et al., 2021; Poria et al., 2015, 2016). However, with the increase in the number of modalities, the processing becomes more and more difficult. The trickiest problem that needs to solve is fusing different modalities efficiently. Generally, fusion methods can be categorized into manners, one is early fusion and the other is late fusion (Abdu et al., 2021). Early fusion simply concatenates the features of different modalities as the input. Late fusion builds separate models for each modality and then integrates the outputs. As these methods cannot exploit the interactions between modalities, more efficient networks are proposed such as TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MFN (Zadeh, Liang, Mazumder et al., 2018), DFG (Zadeh, Liang, Poria et al., 2018), MULT (Tsai et al., 2019), MFRM (Mai et al., 2022). Besides these, hierarchical fusion is also widely studied in which bimodal interactions are generated based on unimodal information, and trimodal interactions are obtained based on bimodal information. They are end-to-end or sequences to sequences or step by step (Gu et al., 2018; Han et al., 2021; Huddar et al., 2020; Mai et al., 2020; Majumder et al., 2018; Poria et al., 2017a, 2017b; Tang et al., 2021).

To focus on the important features, attention-based methods are put forward the multi-modal multi-utterance attention framework for sentiment prediction (Ghosal et al., 2018). Chauhan et al. (2019) proposed the context-aware interactive attention for multimodal sentiment and emotion analysis, which learns the inter-modal interaction among various modalities and exploits the correspondence among the neighboring utterance. Huddar et al. (2020) calculated the bimodal attention matrix representations separately, and concatenated them as the trimodal attention matrix to fuse the interaction information from different modalities. Xi et al. (2020) proposed a method based on the multi-head attention mechanism, which uses the self-attention mechanism to extract the intra-modal features and employs the multi-head mutual

Table 1
Datasets statistics in CH-SIMS, CMU-MOSI, MOSEI.

Dataset	Train	Valid	Test	Total
CH-SIMS	1368	456	457	2281
CMU-MOSI	1284	229	686	2199
MOSEI	16 326	1871	4659	22 856

attention to analyze the correlation between different modalities. Wu et al. (2022) designed a bimodal information-augmented multi-head attention to fuse the information carried by different modalities and explore the relative importance of pair-wise modalities.

In addition to these fusion strategies and attention-based methods, MTL is another popular method. It takes the analysis of multiple modalities as multiple tasks via hard or soft parameter sharing. In hard parameter sharing, a subset of the parameters is shared between tasks while other parameters are task-specific. In soft parameter sharing, all parameters are task-specific but they are jointly constrained by some principles (Sener & Koltun, 2018). Hard parameters sharing-based MTL is more popular in the previous studies for multimodal sentiment analysis. As an effective learning paradigm, Tian et al. (2018) applied it to multimodal sentiment analysis, with sentiment score prediction as the main task and polarity and/or intensity classification as the auxiliary tasks. Akhtar et al. (2019) proposed a deep multi-task learning framework that jointly performs sentiment and emotion analysis. Based on this work, Akhtar et al. (2020) further exploited the relatedness among the participating tasks (including sentiment & emotion classification, sentiment classification & sentiment intensity prediction, emotion classification & emotion intensity prediction) and leveraged them for the overall performance improvement. Chauhan et al. (2020b) proposed the multi-task framework which uses sentiment and emotion to help predict sarcasm. Besides, Chauhan et al. (2020a) explored the relationships and similarities of a variety of tasks including humour, sarcasm, offensive, motivation, and sentiment detection. Fortin and Chaib-draa (2019) applied MTL in image-text sentiment analysis to address the modality-missing cases. To solve the time-consuming and labor cost problem of unimodal annotations, Yu et al. (2021) constructed a self-supervised multi-task multimodal sentiment analysis network(Self-MM) where the unimodal labels are obtained by the model via multimodal labels. A two-phase multi-task framework was introduced by Yang et al. (2022) for multimodal sentiment analysis. They first fine-tuned BERT to learn the contextual text representations and then explored cross-modal attention between text modality and other modalities for multi-task learning. An ensemble of models is a set of learning models whose individual predictions are combined in some way to attain a more generalizable result. Decision-level fusion can be regarded as a method of ensemble learning. Poria, Peng et al. (2017) proposed a framework which integrates convolutional neural networks and multiple kernel learning. However, most ensemble learning based approaches still aim at textual tasks (Gaye et al., 2021; Ye et al., 2021).

3. Datasets

To facilitate the exposition of the main motivations and the proposed model in our paper, we first introduce extension of the datasets. We use three public multimodal sentiment analysis datasets, CH-SIMS (Yu, Xu et al., 2020), CMU-MOSI (Zadeh et al., 2016), MOSEI (Zadeh, Liang, Poria et al., 2018). The first one is in Chinese, and the others are in English. The statistics about the datasets are shown in Table 1.

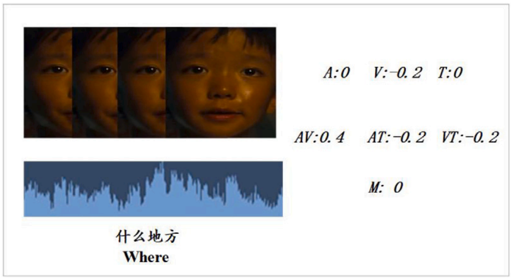

Fig. 1. A sample from CH-SIMS with fine-grained labels, A, V, T denote unimodal labels, AV, AT, VT denote bimodal labels. M means multimodal label.

Table 2
Different hierarchical sentiment distribution in CH-SIMS. For example, 502 samples are Negative with respect to AV.

Modal	NG	WN	NU	WP	PS
A	698	499	367	483	234
T	761	436	338	290	456
V	611	502	520	366	282
AV	502	744	254	441	340
AT	626	639	169	542	305
VT	591	706	222	385	377
M	754	484	345	346	352

CH-SIMS (Yu, Xu et al., 2020) contains 60 raw videos and 2281 refined video segments. The length of the clips is no less than one second and no more than ten seconds. Besides, this dataset has already provided both unimodal labels and multimodal labels with manual sentiment annotation in the range [−1, 1]. Research is constrained to explore the contribution of different hierarchical modalities to multimodal sentiment analysis, as bimodal labels are unavailable. The multimodal sentiment datasets mainly contain unimodal data such as text, audio, and video. Therefore, we treat the two modalities as bimodal data: text-audio, audio-video, and video-text. In this paper, to mine the impact of more fine-grained modal labels on multimodal sentiment, each clip with bimodal granularity is annotated CH-SIMS as bimodal labels.

Specifically, we invite 15 students (including undergraduate and postgraduate) and divide them into three groups. Each group has 5 students and makes five labels for each utterance. They annotate the sentiment with the specific values of −1 (negative), 0 (neutral), 1 (positive). Besides, the first group can only judge the sentiment with audio and text to form AT labels. The second group decides by combining text and silent video (VT labels). The third group utilizes the audio and silent video to help annotate (AV labels). After getting the five labels of AV, AT, VT of each utterance respectively, we average them to get the final sentiment label. In other words, the final labeling results are one of {−1.0, −0.8, −0.6, −0.4, −0.2, 0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. These values are further divided into 5 categories: Negative (NG) {−1.0, −0.8}, Weakly Negative (WN) {−0.6, −0.4, −0.2}, Neutral (NU) {0}, Weakly Positive (WP) {0.2, 0.4, 0.6}, Positive (PS) {0.8, 1.0}. Fig. 1 shows an example from CH-SIMS with different hierarchical sentiment labels. The statistics of unimodal sentiment (A, V, T), bimodal sentiment (AV, AT, VT), and multimodal sentiment (M) categories in CH-SIMS are demonstrated in Table 2.

In this dataset, the training, validation, and test set are in proportion 6:2:2. Furthermore, the sentiment distributions of the split dataset are presented in Fig. 2, where Fig. 2(a), Fig. 2(b), Fig. 2(c) shows the sentiment distribution of the training set, the validation set and the test set respectively. For the legend in the figure, A, V, T represent the unimodal labels, AV, AT, VT represent the bimodal labels and M represents the multimodal labels. Besides, the ordinate means the number of samples of the corresponding modality with a specific category. We can see that the sentiments of different modalities are different. Among them, CH-SIMS has 113 video clips with different sentiment polarities for each modality. The modalities have potential influences on each other.

We choose the English dataset CMU-MOSI for expansion to enrich the study and subsequent experimental verification in our paper. CMU-MOSI contains 93 opinion videos collected from YouTube movie reviews that were created by Zadeh et al. (2016). Each video is split

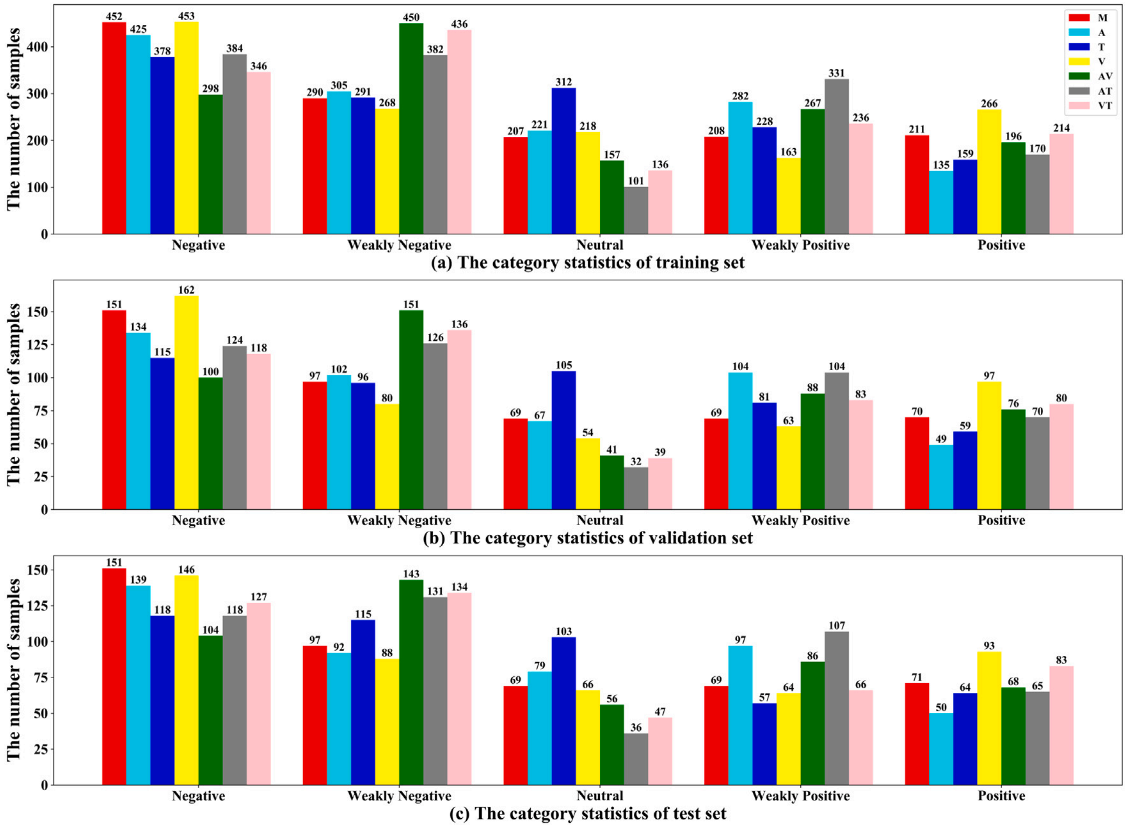

Fig. 2. The split of CH-SIMS and sentiment distribution. (a) The sentiment distribution of the training set. (b) The sentiment distribution of the validation set. (c) The sentiment distribution of the test set.

into short segments. The final dataset consists of 2199 short monologue video clips with manual sentiment annotation in the range [−3, 3], which corresponds to the sentiment score ranging from highly negative to highly positive. The CMU-MOSI also requires unimodal sentiment annotation as unimodal labels, unlike CH-SIMS dataset. Therefore, we make three unimodal annotations and three bimodal annotations for each video segment as the CMU-MOSI only contains multimodal labels. The annotating standard of the bimodal sentiment is the same to that CH-SIMS. Specifically, we invite 5 students (postgraduate) to make 5 labels for each modality of each video segment. The students have 8 choices for each video segment: strongly positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (−1), negative (−2), strongly negative (−3). To reduce the influence of the correlation information between modalities on the annotator, we stipulate that the order of annotation is text first, audio second, and silent video last. After getting the five labels of A, V, T of each utterance respectively, we average them to get the final sentiment label. The processed dataset and regenerated features are available.

MOSEI is a large dataset that contains 23 453 annotated video segments from 1000 distinct speakers with 250 topics that were created by Zadeh et al. (2016). Like MOSI, each video segment is annotated manually with sentiment value in the range [−3, 3], and the value refers to one sentiment category from strongly negative to strongly positive.

4. Fine-grained modal label-based multi-stage network

To efficiently utilize the sentiment differences between various modalities and explore the relationship between different modalities, FmlMSN is proposed. It consists of three layers, as illustrated in Fig. 3. The first layer is the unimodal features extraction layer, the second is the multi-stage layer consisting of three base learners of pair-wise modalities trained by multi-task learning. The third is the multimodal fusion and prediction layer. Specifically, in the first layer, the model takes the modalities in each utterance as input and extracts the unimodal features using the feature extraction subnets. Note that V in this paper means an image sequence sampled from the related utterance, and the visual features mentioned in the following sections denote facial features. The multi-stage layer includes three bimodal sentiment predictors, the first for acoustic-visual, the second for acoustic-textual, and the third for visual-textual. In each stage, bimodal sentiment analysis works as the main task and unimodal sentiment analysis works as the auxiliary task. The qualified bimodal interaction features obtained from the three stages serve as the input of the meta-learner for fusion and sentiment prediction. In the training procedure, the training and computation of each bimodal sentiment predictors and multimodal module are independent.

4.1. *Unimodal features extraction*

The utterances of each video act as the basic unit, and N (N is a positive integer) is the number of utterances in the dataset. Then the dataset can be formally represented as $U = \{u_1, u_2, \dots, u_N\}$. The feature extraction methods for all modalities in each utterance are as follows.

4.1.1. *Textual features*

Considering the impacts of inaccurate Chinese word segmentation tools on semantic relationship extraction and the superior performance of BERT in NLP, pre-trained Chinese BERT (Devlin et al., 2019; Xu et al., 2023) is employed to get the sentence embedding of each utterance. Let t_i be a sentence embedding and $t_i = \{w_1, w_2, \dots, w_{L_i}, w_l \in \mathbb{R}^{d_t}\} \in \mathbb{R}^{L_i \times d_t}$, where w_l represents the word embedding with dimension d_t, L_i is the number of words in the sentence. Since the number of words in each sentence is different, padding and truncation are

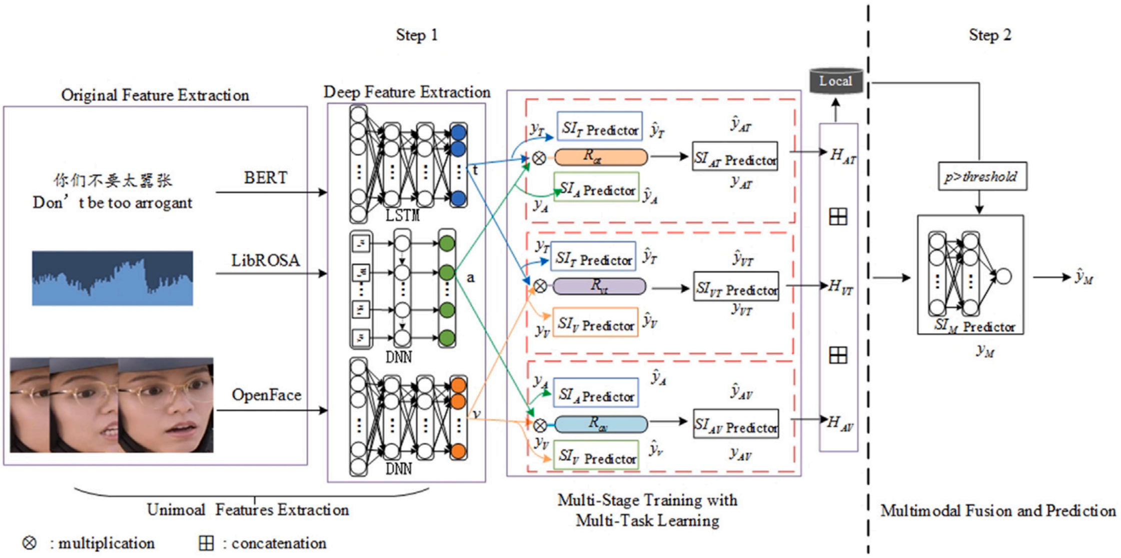

Fig. 3. The network architecture of FmlMSN. It consists of three components: Unimodal Features Extraction; Multi-Stage Training with Multi-Task Learning; Multimodal Fusion and Prediction. SI predictor is to predict the sentiment intensity. Each red dotted box means a training stage. y_A, y_T, y_V are unimodal true labels. y_{AT}, y_{VT}, y_{AV} are bimodal true labels, y_M means multimodal true labels. The hat of y means the corresponding predicted labels.

introduced to let the final length be L. L is calculated in two steps. First, we obtain the average length of the sentences and calculate the standard deviation of raw lengths. Then we take the sum of the average and λ times the standard deviation as the final length. The calculations are demonstrated in Eqs. (1) and (2). In the process of fixing the length, the short sentence is filled with $pad \in \mathbb{R}^{(L-L_i)\times d_t}$ at the end. For the long sentence, the first L vectors are taken to constitute the sentence embedding. As a result, the dimension of t_i is $L \times d_t$.

$$\tilde{L} = \frac{1}{N}\sum_{i=1}^{n} L_i \tag{1}$$

$$L = \tilde{L} + \lambda \times \sqrt{\frac{1}{N}\sum_{i=1}^{N}(L_i - \tilde{L})^2} \tag{2}$$

As there exist complex contextual relationships in word sequences, it is necessary to capture the long-term dependencies between words. To solve this problem, t_i is fed to LSTM to obtain the informative features. We employ the final hidden state output as the sentence embedding with d_{t1} dimension.

4.1.2. Acoustic and visual features

There are many features in audio that can reflect emotions and tone of speech such as Zero Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCCs), and Constant-Q chromatogram (CQT). To obtain such sentimental information, LibROSA (McFee et al., 2015) speech toolkit is used to extract d_a-dimensional frame-level acoustic features at 22 050 Hz, including z-dimensional ZCR, m-dimensional MFCCs and c-dimensional CQT. Similarly, each video is framed at 30 Hz by using multimedia processing toolkit ffmpeg.[1] The aligned faces are extracted via the MTCNN face detection algorithm (Zhang et al., 2016). To obtain the facial features related to emotion, MultiComp OpenFace2.0 toolkit[2] is used to extract d_v-dimensional facial features, including facial landmarks, facial action units, head pose, head orientation, and eye gaze. As we all know, audios and videos are time-series data and contain different numbers of frames. Similar to the processing of text, each audio and video is processed to obtain L' frames and L'' frames, respectively.

To further extract the acoustic and visual features, we average all the features of the frames in each utterance and input them into the three-layer deep neural networks respectively, which consists of three fully connected layers. For audio, each layer has d_{a1} ReLU activation units. For video, each layer has d_{v1} ReLU activation units. The final acoustic features and visual features are $a_i \in \mathbb{R}^{d_{a1}}$, $v_i \in \mathbb{R}^{d_{v1}}$ respectively, where i means the ith utterance.

4.2. Multi-stage training with multi-task learning

As the textual, acoustic and visual features of each utterance are extracted in Section 4.1, the three modalities of the whole dataset can be represented as $R_T = \{t_1, t_2, \ldots, t_N\}$, $R_A = \{a_1, a_2, \ldots, a_N\}$, $R_V = \{v_1, v_2, \ldots, v_N\}$, where $R_T \in \mathbb{R}^{d_{t1}}$, $R_A \in \mathbb{R}^{d_{a1}}$, $R_V \in \mathbb{R}^{d_{v1}}$. Different from the previous multi-task methods, the novel idea of the proposed model is to perform pair-wise multi-task learning with specific label of modal granularity in each stage. Assume that α and β are any two modalities of the three, γ denotes the combination of them. α, β, γ corresponds to three different tasks respectively, where $\alpha, \beta \in \{A, V, T\}$, $\gamma \in \{AV, AT, VT\}$. In each task, the goal is to predict the sentiment intensity (SI), and essentially it is a regression mission. The multi-task learning based on pair-wise modalities is shown in Fig. 4, where SI_α, SI_β, SI_γ is the sentiment intensity predictor of α, β, γ respectively. The calculation of R_γ is governed by Eq. (3), which is actually the outer product from R_α and R_β. For example, let α be A, β be T, the dimension of R_{AT} is $d_{a1} \times d_{t1}$.

$$R_\gamma = R_\alpha \otimes R_\beta \tag{3}$$

In this training process, each task shares the unimodal features extraction layer. However, each one has a separate predictor with a three-layer deep neural network structure. For unimodal tasks, the predictors have d_{im} $(i \in \{t, a, v\})$ units in the hidden layers and one output. For bimodal tasks, the predictors have d_m units in the hidden layers and one output. Eq. (4) represents the SI predictor predicting unimodal sentiment, and Eq. (5) represents the SI predictor predicting bimodal sentiment.

$$\begin{aligned} \hat{y}_\alpha &= SI_\alpha(R_\alpha) \\ \hat{y}_\beta &= SI_\beta(R_\beta) \end{aligned} \tag{4}$$

[1] https://ffmpeg.org/ffmpeg.html.

[2] https://github.com/TadasBaltrusaitis/OpenFace/wiki.

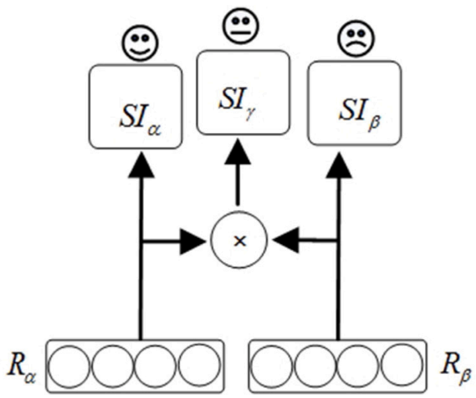

Fig. 4. Two-two modalities based multi-task learning in each stage.

$$\hat{y}_\gamma = SI_\gamma(R_\gamma) \tag{5}$$

Besides, for different tasks, the optimization objective of training is:

$$min\frac{1}{N}\sum_{i=1}^{N}(\omega_\alpha\phi(\hat{y}_{\alpha,i}, y_{\alpha,i}) + \omega_\beta\phi(\hat{y}_{\beta,i}, y_{\beta,i}) + \omega_\gamma\phi(\hat{y}_{\gamma,i}, y_{\gamma,i})) \tag{6}$$

where y_α, y_β, and y_γ are the true labels, N is the number of training samples, ω_α, ω_β, and ω_γ are the hyperparameters to balance different tasks. $\phi(\hat{y}, y)$ means the training loss function which is the L1 loss in this paper. Note that the y_α, y_β, y_γ, and y_M are different in this paper which corresponds to unimodal labels, bimodal labels, and multimodal labels respectively. Since our work is essentially a regression mission for predicting sentiment intensity, the L1 loss is used.

4.3. Multimodal fusion and prediction

To better and efficiently fuse multimodal information, we save the activations of the second layer of SI_{AV}, SI_{AT} and SI_{VT} respectively, which are represented as H_{AV}, H_{AT}, H_{VT}. They serve as the input of multimodal predictor SI_M, which is a three-layer deep neural network with d_m units in the hidden layer and one output. Eq. (7) shows the formal expression for the above descriptions, where F means the fusion function. In this paper, F uses concatenation.

$$\hat{y}_M = SI_M(F(H_{AV}, H_{AT}, H_{VT})) \tag{7}$$

In this component, the L1 loss acts as the training loss. In addition, it is worth noting that different from the previous end-to-end methods, the multimodal module is an independent part with some alternative settings on the input, namely the activations can also be saved locally and utilized as the information of multimodal fusion depending on the situation, rather than training from scratch. In this way, the model can achieve modularization and save a lot of time and computing resources.

The whole training procedure is shown in Algorithm 1. In the procedure, getUnimodalFeature(U) is to vectorize the textual, acoustic, and visual features in the utterances. It extracts the unimodal original features and the deep features. The extraction methods have been explained in Section 4.1 concretely. In TRAIN(R_T, R_A, R_V), there mainly contain two steps, one is multi-stage training, the other is multimodal module for multimodal fusion and prediction. For the multi-stage training, the extracted features are fed to the two-two modalities based multi-task learning network. For the multimodal module, an alternative setting is provided before fusing the multiple modalities. Judge whether it needs to retrain the bimodal predictors or not. If so, step 1 is executed first followed by step 2. Otherwise, the locally preserved features are directly fused for the final sentiment prediction. In other words, only step 2 is performed. The details of step 1 and step 2 can refer to Section 4.2 and Section 4.3. The criterion of when to retrain depends on the validity of the model on the test cases. The formula is shown in Eq. (8), where TP denotes the number of truly predicted samples, $Total$ presents the number of test samples that are randomly selected from the whole validation dataset. $\frac{TP}{Total}$ means the proportion of the test samples that are correctly predicted. We calculate this ratio M times. p is the average of these ratios. If p is lower than the threshold, then step 1 is executed. The specific operation refers to Section 5.5.

$$p = \frac{1}{M}\sum_{i=1}^{M}(\frac{TP}{Total})_i \tag{8}$$

Algorithm 1 The Training of FmlMSN

```
procedure getUnimodalFeatures($U$)
# Unimodal Original Features Extraction
for $i$ : $[1, N]$ do
    $t_i \leftarrow getTextEmbedding(u_i)$
    $a_i \leftarrow getAudioEmbedding(u_i)$
    $v_i \leftarrow getVideoEmbedding(u_i)$
end for
# Unimodal Deep Features Extraction
for $i$ : $[1, N]$ do
    $t_i \leftarrow textSubnet(t_i)$
    $a_i \leftarrow audioSubnet(a_i)$
    $v_i \leftarrow videoSubnet(v_i)$
end for
procedure TRAIN($R_T$, $R_A$, $R_V$)
if run multi-stage learning then
    perform step 1;
    use bimodal features obtained from step 1, then perform step 2.
else
    load the locally saved bimodal features then skip straight to step
    2.
end if
step1: # Pair-wise Modalities based Multi-Task Learning
    $H_{AV}, \hat{y}_A, \hat{y}_V, \hat{y}_{AV} \leftarrow train\_AV(R_A, R_V)$
    $H_{AT}, \hat{y}_A, \hat{y}_T, \hat{y}_{AT} \leftarrow train\_AT(R_A, R_T)$
    $H_{VT}, \hat{y}_V, \hat{y}_T, \hat{y}_{VT} \leftarrow train\_VT(R_V, R_T)$
    save $H_{AV}, H_{AT}, H_{VT}$.
step2: # Multimodal Fusion and Prediction
    $R_M \leftarrow H_{AV} \oplus H_{AT} \oplus H_{VT}$
    $\hat{y}_M \leftarrow SI_M(R_M)$
    return $\hat{y}_M$
```

5. Experiments and results analysis

5.1. Experimental settings

In the training procedure, we consider tuning the following hyperparameters: learning rate (lr), batch size (bs), dropout (tdrp, adrp, vdrp) and number of hidden units of each modality-specific subnetwork (thid, ahid, vhid), out dimensions of text subnetwork (tout), hidden units of unimodal predictors (tdim, adim, vdim), dropout and units of bimodal and multimodal prediction layers (fdrp, fdim). The values for each hyperparameter of different datasets are shown in Table 3.

We use Adam optimizer with initial learning rate throughout all experiments and perform early stopping by 20 epochs. Besides, in Section 4, the values of d_t, d_a, d_v are 768, 33, 709; the values of L, L', L'' are 39, 400, 55; d_{t1}, d_{a1}, and d_{v1} refer to tout, ahid and vhid; d_{am}, d_{vm}, d_{tm} refer to tdim, adim, vdim. d_m corresponds to fdim; λ is 3; the padding character in pad is 0; z is 1, m is 20 and c is 12; ω_α, ω_β, ω_γ are 1. The same as that in literature (Yu, Xu et al., 2020), we record the experimental results on CH-SIMS in two forms: multi-class classification and regression. For multi-class classification, we report 2-class accuracy (Acc-2), 3-class accuracy (Acc-3), 5-class accuracy (Acc-5), and Weighted F1 score (F1). For regression, we report Mean Absolute Error (MAE), and Pearson Correlation (Corr). For MOSI

Table 3
Hyperparameters of FmlMSN. For MOSI and MOSEI, the -a denotes the parameters for the aligned dataset, while the -na denotes the parameters for the unaligned dataset.

Para	Dataset				
	CH-SIMS	MOSI-a	MOSI-na	MOSEI-a	MOSEI-na
lr	0.001	0.001	0.002	0.001	0.002
bs	64	64	64	64	64
tdrp	0	0.1	0.2	0.1	0.1
adrp	0	0.1	0.2	0.1	0.1
vdrp	0	0.1	0.2	0.1	0.1
thid	128	128	128	128	64
ahid	16	32	32	16	64
vhid	128	64	64	64	64
tout	128	32	128	64	64
fdrp	0	0	0.1	0	0
fdim	64	16	32	64	16
tdim	64	16	32	8	32
adim	16	16	64	32	32
vdim	64	64	64	8	16

and MOSEI dataset, the metrics are 7-class accuracy (Acc-7), Acc-5, 2-class accuracy, Weighted F1 score, MAE, and Corr. Different from the Acc-2 and F1 of CH-SIMS, following (Yu et al., 2021), the Acc-2 and F1-score are calculated in two ways: negative/non-negative (non-exclude zero) and negative/positive (exclude zero). For all metrics mentioned, except for MAE, the higher value means the better.

5.2. Comparisons of different models

EF-LSTM. The EF-LSTM (Williams et al., 2018) concatenates the original features of the three modalities and inputs them into the LSTM to capture the long-term dependencies between the modal sequences.

LF-DNN. The LF-DNN (Yu, Xu et al., 2020) uses the DNN to learn unimodal features and then concatenates them as the input of the prediction layer.

TFN. The TFN (Zadeh et al., 2017) captures multimodal interaction information by creating a multi-dimensional tensor.

LMF. The LMF (Liu et al., 2018) is the improvement over TFN, where the low-rank multimodal tensors fusion technique is adopted to improve the efficiency.

MFN. The MFN (Zadeh, Liang, Mazumder et al., 2018) stores the internal information of the modalities and the interaction information between the modalities through the gated memory unit. It adds dynamic fusion graphs to reflect effective emotional information.

DFG. The DFG (Zadeh, Liang, Poria et al., 2018) replaces the fusion block in MFN with a Dynamic Fusion Graph, which is directly related to how modalities interact.

MulT. The MulT (Tsai et al., 2019) uses its cross-modal attention module to extract the key information inside each modality and then merges these features based on the Transformer (Tsai et al., 2019) model.

MISA. The MISA (Hazarika et al., 2020) incorporates the combination of losses including distributional similarity, orthogonal loss, reconstruction loss, and task prediction loss to learn modality-invariant and modality-specific representation.

MLF-DNN, MTFN, MLMF. These are multi-task frameworks of the LF-DNN, TFN, and LMF. All of them use independent unimodal labels (Yu, Xu et al., 2020).

RAVEN. The RAVEN (Wang et al., 2019) uses nonverbal behaviors to dynamically model linguistic features. It captures the context intention by shifting work representation.

MCTN. The MCTN (Pham et al., 2019) inputs a target modality data to learn joint representations in the training phase. It uses the sequence to sequence model and machine translation.

bc-LSTM. The bc-LSTM (Poria et al., 2015) mines contextual information based on the LSTM model.

Table 4
Results of different models for sentiment analysis.

Model	Acc-2	Acc-3	Acc-5	F1	MAE	Corr
EF-LSTM	69.37	54.27	21.23	56.82	0.591	0.055
LF-DNN	77.02	64.33	39.74	77.27	0.446	0.555
TFN	78.38	65.12	39.30	78.62	0.432	0.591
LMF	77.77	64.68	40.53	77.88	0.441	0.576
MFN	77.90	65.73	39.47	77.88	0.435	0.582
DFG	78.77	65.65	39.82	78.21	0.445	0.578
MulT	78.56	64.77	37.94	79.66	0.453	0.564
MISA	69.37	51.42	20.79	56.82	0.587	0.113
MLF-DNN	80.44	69.37	40.22	80.28	0.396	0.665
MTFN	81.09	68.80	40.31	81.01	0.395	0.666
MLMF	79.34	68.36	41.05	79.07	0.409	0.639
Self-MM	80.04	65.47	41.53	80.44	0.425	0.595
FmlMSN	**83.59**	**72.87**	**48.14**	**83.64**	**0.377**	**0.690**

FAF. The FAF (Gu et al., 2018) explores the relationship between text and audio of each word based on word-level fusion.

MMMU-BA. The MMMU-BA (Ghosal et al., 2018) employs the Bidirectional Gate Recurrent Unit (Bi-GRU) and attention Model.

CIA. The CIA (Chauhan et al., 2019) learns the interactive information of pair-wise modalities through an auto-encoder mechanism.

MARN. The MARN (Amir et al., 2018) uses multi-attention Block and Long-short Term Hybrid Memory(LSTHM).

MTL. The MTL (Akhtar et al., 2019) is a multi-task learning framework. It uses pair-wise inter-modal attention mechanism to learn information.

CTC. The CTC (Graves et al., 2006) employs Recurrent neural network (RNN) to label sequence data.

Self-MM. The Self-MM (Yu et al., 2021) designs a unimodal label generation strategy based on the self-supervised method. It introduces unimodal subtasks to aid in learning modality-specific representations.

In CH-SIMS, some video utterances contain a lot of noises. For example, an utterance is with irrelevant video clips, the same utterance contains multiple figures, the text contains typos and the video content does not match the corresponding sentences. To standardize the dataset, we make sure that there is only one person in each utterance. Besides, we correct the typos and the mismatches. Based on the improved dataset, we re-extract the features. The processed dataset and regenerated features are provided.[3]

Consider the results of the previous papers are obtained on the unprocessed dataset, we also conduct experiments on the old version for a fair comparison. The experimental results are shown in Table 4.

From the results in Table 4, it is easy to figure out that the results of multi-task learning based methods (from MLF-DNN to FmlMSN) are better than those of other end-to-end fusion methods on most metrics. In particular, FmlMSN has better performance on all evaluation metrics. It achieves dominant results, especially in five-class and regression tasks. The reasons may be that FmlMSN has fully extracted the unimodal features and obtained the bimodal features containing rich affective information with fine-grained modal labels. Besides, MLF-DNN, MTFN, MLMF, and Self-MM are multi-task learning-based methods whose tasks include T, A, V, and M. Compared with them, FmlMSN also considers the bimodal tasks, namely AV, AT, and VT. As we can see, FmlMSN achieves better performance than other models do. It uses unimodal tasks as auxiliaries and bimodal tasks as the main task to obtain the bimodal features, which can better serve as the input of the multimodal module.

To further evaluate the effectiveness of the proposed model and observe the effects of the above noises, we compare our model with the above methods by using the new features extracted from the processed dataset. The results are shown in Table 5.

[3] https://drive.google.com/drive/folders/1Abl6tj2L9cTw2pHEI6A1casXegSR8XSz?usp=sharing.

Table 5
Results of experiments on new features. Significance T-test (<0.05) signifies that the obtained results are statistically significant over the existing methods with 95% confidence score.

Model	Acc-2	Acc-3	Acc-5	F1	MAE	Corr
EF-LSTM	69.37	54.27	21.23	81.91	0.590	0.035
LF-DNN	79.65	67.40	**44.64**	80.04	0.419	0.591
TFN	80.09	65.86	41.79	80.69	0.408	0.613
LMF	79.43	66.52	42.89	79.75	0.411	0.630
MFN	77.90	63.89	40.04	78.27	0.447	0.564
DFG	75.93	62.80	37.20	76.13	0.448	0.556
MulT	77.46	64.33	34.57	78.27	0.448	0.563
MISA	79.21	61.27	36.32	78.86	0.438	0.569
MLF-DNN	83.37	69.15	40.70	**83.80**	0.399	0.678
MTFN	81.84	65.43	37.86	81.63	0.426	0.601
MLMF	81.40	69.15	37.20	81.42	0.414	0.649
Self-MM	77.90	63.46	42.01	78.05	0.424	0.569
FmlMSN	**83.59**	**70.90**	41.79	83.71	**0.385**	**0.680**
T-test	0.0192	0.0082	0.0209	0.0386	0.0044	0.0143

The experimental results show that the new features have some effects on the above methods. EF-LSTM obtains great improvement on F1 but drop on Corr. LF-DNN, TFN, LMF, MISA perform better on most metrics, especially in regression task. However, the performance of MFN, DFG, MulT decreases to some degree. For the multi-task learning-based methods, except MLF-DNN, the results of others are not as good as those in Table 4. The common phenomenon is the decrease in MAE or Corr. This may in turn affect other metrics. Essentially, our model tends to predict the sentiment intensity, which is a regression task. The evaluations of the classification task are based on the results of the regression results. For example, Acc-2 is calculated based on the values greater than 0 as positive and less than 0 as negative. Acc-3 and Acc-5 are obtained in the same way. On the whole, the proposed method is better than the mentioned four multi-task learning and end-to-end methods. However, we note that in general there is a performance drop on FmlMSN when we replace the features extracted from the raw dataset with the new features, especially on Acc-5. The reasons are considered from two aspects. One is the influence of the performance drop of regression, the other is due to the data processing. In fact, we crop the video to filter out the extra faces and replace the incorrect video (video that already exists in the dataset or mismatches with the transcript) with the new one. The operations not only reduce the facial features but also affect the raw sentiment labels.

We perform statistical significance test (paired T-test) on the obtained results. As reported in Table 5, we observe that the performance achieved in our proposed approach is significantly better in comparison to the state-of-the-art methods with p-value < 0.05.

5.3. Importance of the fine-grained labels

The fine-grained sentiment labels (unimodal, bimodal, and multimodal) are introduced in this work. To confirm the impact of using different labels of modal granularity on multimodal sentiment analysis, the bimodal labels and the multimodal labels are applied to the FmlMSN model respectively (corresponding to FmlMSN and FmlMSN-m in Table 6). Note that except for the experiments in Table 4, all the experiments are performed with new features. The results of FmlMSN and FmlMSN-m are shown in Table 6. As we can see, the performance with bimodal information is better than that with multimodal information. The result is expected since Fig. 4 has already given us much intuitional insight into the sentiment differences between different modalities. These differences directly affect the sentiment analysis of different modalities. Therefore, for bimodal features extraction and fusion, using bimodal labels may be a good choice.

Table 6
The experimental results of FmlMSN with/without fine-grained labels.

Model	Acc-2	Acc-3	Acc-5	F1	MAE	Corr
FmlMSN	**83.59**	**70.90**	**41.79**	**83.71**	**0.385**	**0.680**
FmlMSN-m	81.40	64.77	40.92	82.30	0.418	0.636

Table 7
Results for multimodal sentiment analysis with different tasks. T, A, V represents text, audio, video task, respectively. Bimodal labels are used for the bimodal task.

Modality		Acc-2	Acc-3	Acc-5	F1	MAE	Corr
A, V	A	67.83	50.55	20.13	80.83	0.540	0.116
	V	75.93	62.36	32.82	76.00	0.493	0.559
	$A+V$	74.18	59.96	31.95	74.72	0.467	0.459
A, T	A	64.77	44.64	25.82	68.90	0.501	0.270
	T	73.52	69.15	50.98	72.07	0.318	0.691
	$A+T$	78.12	70.90	43.98	78.39	0.407	0.594
V, T	V	80.96	66.52	38.51	81.73	0.437	0.619
	T	85.12	71.99	51.86	85.08	0.309	0.730
	$V+T$	82.06	73.30	46.83	82.19	0.351	0.691
M	$A+V+T$	83.59	70.90	41.79	83.71	0.385	0.680

Table 8
Results for multimodal sentiment analysis with different tasks. Multimodal labels is used for the bimodal tasks.

Modality		Acc-2	Acc-3	Acc-5	F1	MAE	Corr
A, V	A	67.83	50.55	21.66	80.83	0.536	0.102
	V	79.65	66.08	35.67	80.53	0.465	0.596
	$A+V$	75.27	63.02	34.13	76.24	0.480	0.494
A, T	A	63.89	47.70	24.73	69.47	0.514	0.216
	T	83.81	67.40	46.61	83.35	0.329	0.695
	$A+T$	76.59	64.77	41.36	76.52	0.437	0.565
V, T	V	79.21	66.30	40.48	79.44	0.433	0.609
	T	84.46	65.21	45.95	84.81	0.350	0.685
	$V+T$	80.96	68.49	45.08	81.43	0.403	0.641
M	$A+V+T$	81.40	64.77	40.92	82.30	0.418	0.636

5.4. Ablation study

To further observe the performance of the model and the importance of different modalities, we report the unimodal, bimodal, and multimodal results in Tables 7 and 8. Table 7 shows the results of the model using bimodal labels in the bimodal multi-task learning module. Table 8 reports the results of the model using multimodal labels in each stage. Note that each stage is a group containing three tasks, i.e., unimodal tasks and bimodal task. We demonstrate the results of all tasks and the multimodal module.

From the results in Tables 7 and 8, in the unimodal tasks, we can see that text modality (T) generally provides more important knowledge than video modality (V) and audio modality (A) do. The model with V achieves the best performance without T. T is the best because the text provides more semantic information than other modalities. A performs the worst. The reason may be that the short duration and noises in some audios, which affect the quality of the extracted acoustic features. In the bimodal tasks, we observe that VT is the best, followed by AT, and AV. The excellent performance of VT and AT may owe to the assist of text. However, in each stage, the bimodal task does not always outperform the unimodal tasks but is at least better than one. For example, in AV from Table 7, the combining of A and V is better than the single A but not as good as V. In VT from Table 8, the performance of T is the best, the combining of A and V is the second, and V is the worst. In addition, the performance of M has been improved based on these unimodal and bimodal tasks. It outperforms VT on Acc-2 and F1, but it does not perform as well on other metrics. The results seem to show that trimodal analysis may not always help. However, as the sentiment differences between various modalities, the effects of different combinations of modalities are also different. The experiments

Table 9
Comparison between two-step training and local features based training.

Model	Acc-2	Acc-3	Acc-5	F1	MAE	Corr	Time (s)
FmlMSN	83.59	**70.90**	41.79	**83.71**	0.385	0.680	4.8970
FmlMSN-L	83.59	70.68	41.79	83.57	0.385	0.680	**0.8787**

of Table 7 consider these differences and use fine-grained labels as true labels.

For comparison, the experiments of Table 8 only use the multimodal labels as the true labels in the bimodal tasks and multimodal module. From Table 8, in acoustic-visual multi-task stage, the F1 of V is marginally lower than that of A (0.30%), while all other indicators achieve the best performance. In acoustic-textual multi-task stage, T gets overwhelming results. In visual-textual multi-task stage, compared with V, T only performs slightly not as well as it does in Acc-3. The model with T outperforms A on all metrics in Table 7. The reason for the results is that the tasks in Table 7 employ fine-grained labels and the tasks in Table 8 only use multimodal labels. Compared with fine-grained labels, the multimodal labels are not suitable for unimodal tasks and bimodal tasks. Overall, in unimodal tasks, we observe that the performance of T is the best, next V, and least A. As we can see, except for AV, the results of bimodal tasks have decreased to some extent. The M performs not very well. The reason is that multimodal labels cannot reflect all the sentiments in various modalities. Thus it fails to extract more complementary information between modalities.

In the training process, we observe that the AT task always arrives at the best epoch earlier than other tasks. In this sense, the bad results of A and AT are explained from two aspects. One is the noises from raw audio. The other is deficient training due to the same parameter settings for different bimodal tasks. We will fully explore this impact in future.

5.5. Comparison between two-step training and local features based training

It is a two-step process because we first carry out the multi-task training for two-two modalities, and then carry out the multimodal fusion and sentiment prediction. To observe the efficiency of such architecture, we set whether to carry out multi-task learning or not, that is, to conduct step by step (FmlMSN) or to perform the second step directly using locally saved bimodal features (FmlMSN-L). The experimental results are shown in Table 9. In general, FmlMSN is better than FmlMSN-L. However, for the time cost of an epoch under the experimental environment of this paper, FmlMSN-L is about 1/6 of that of FmlMSN, it is very time-saving because it does not need to train the bimodal predictors again.

As for when to retrain the bimodal multitask frameworks, it depends on the validity of the model on the test cases. Specifically, when we first obtain the bimodal features and pre-trained model, we apply them to test for five times. Each test contains ten samples randomly selected from the validation set. We calculate the accuracy of each test and average them to get the p. If p is lower than 74%, retraining is required. The formulas can refer to Eq. (8). Note that the threshold 74% is the lowest result obtained after more than forty sets of tests. The importance of retraining is that if the accuracy of multimodal prediction is not satisfying, the direct reason is the low quality of bimodal features, since they are the input of the multimodal fusion and prediction module. The motivation of this design is to reduce the time consumption without too much impact on the performance of the model. Meanwhile, it separates bimodal and multimodal module for the further optimization and improvement of the model.

Table 10
Results of different models for sentiment analysis on MOSI.

Model	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	DataSetting
bc-LSTM	–	–	80.30	–	–	–	–
FAF	–	–	76.50	76.80	–	–	–
MMMU-BA	–	–	82.31	–	–	–	–
CIA	38.92	–	79.88	79.54	0.914	0.689	–
MTL	–	–	–	–	–	–	–
TFN	32.10	–	73.90	73.40	0.970	0.633	Aligned
EF-LSTM	**35.90**	40.15	78.48	78.51	0.949	0.669	Aligned
MFN	35.83	40.47	78.87	78.90	0.927	0.670	Aligned
DFG	34.64	38.63	78.35	78.35	0.956	0.649	Aligned
MARN	34.70	–	77.10	77.00	0.968	0.625	Aligned
MCTN	35.60	–	79.30	79.10	**0.909**	0.676	Aligned
RAVEN	33.20	–	78.00	76.60	0.915	**0.690**	Aligned
FmlMSN	31.44	**40.61**	**80.09**	**79.78**	0.977	0.669	Aligned

5.6. Experiments on english datasets

To further study the effectiveness and generalization ability of our model, we conduct experiments on the English dataset CMU-MOSI. Considering that there are no fine-grained labels in the CMU-MOSI dataset, we make three unimodal annotations and three bimodal annotations on the English sentiment dataset (CMU-MOSI) to make the FmlMSN work smoothly. Multi-attention Recurrent Network (MARN) (Amir et al., 2018), Recurrent Attended Variation Embedding Network (RAVEN) (Wang et al., 2019), and Multimodal Cyclic Translation Network (MCTN) (Pham et al., 2019) have achieved SOTA results on various word-aligned multimodal tasks for English datasets. Differences in language and structure between Chinese dataset and English dataset, for the sake of fairness, our comparative experiments on the CMU-MOSI dataset deviates from the CH-SIMS dataset.

The results of the proposed FmlMSN model on CMU-MOSI dataset are shown in Table 10. The first five (bc-LSTM, FAF, MMMU-BA, CIA, and MTL) are popular methods based on contextual relation, attention, or multi-task learning. The datasets of these popular methods are obtained from CMU Multi-modal Data SDK[4] or provided in Poria, Peng et al. (2017). The results of the five algorithms does not provide their data setting on CMU-MOSI dataset, we apply the '–' to label them. It is easy to figure out CIA performs better than FAF. In our paper, the datasets are provided in Yu et al. (2021). In addition, we compare the FmlMSN with seven algorithms on word-aligned dataset. We observe that our FmlMSN model achieves competitive performance on Acc-2, Acc-5, and F1 scores. The underlying reason of the results could be that the FmlMSN has considered the unimodal sentiment labels, bimodal sentiment labels, and multimodal sentiment labels. Experimental results indicate that FmlMSN can extract the different hierarchical modalities sentiment feature and make a reasonable trade-off between three bimodal tasks.

The proposed model has a clear superiority on both CH-SIMS and CMU-MOSI with fine-grained modal labels. To further verify the feasibility of the FmlMSN on the MOSEI dataset only with multimodal labels, we perform experiments only use the multimodal labels at all stages including word-aligned and unaligned datasets. For word-aligned dataset, since the MARN was proposed in 2018, the library of the code which is provided by the paper has been deprecated. What is more, the experiment results of MARN are available on the MOSEI dataset. we could not provide the corresponding results and use the '–' mark them in Table 11. We will try to implement it if conditions permit in future. For unaligned datasets, some methods need to use connectionist temporal classification (CTC) (Graves et al., 2006) to adapt it. The results of the MOSEI are shown in Table 11, it is worth noting that FmlMSN-m$_1$ represents the model for aligned dataset, and FmlMSN-m$_2$ means the model for unaligned dataset. From Table 11, we can

[4] https://github.com/A2Zadeh/CMU-MultimodalDataSDK.

Table 11
The experimental results on MOSEI (bold number means the best results of aligned/unaligned datasets. For Acc-2 and F1, the number on the left of '/' denotes "negative/non-negative" and the right is "negative/positive").

Model	MOSEI						Data setting
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	
bc-LSTM	–	–	–	–	–	–	–
FAF	–	–	–	–	–	–	–
MMMU-BA	–	–	79.80	–	–	–	–
CIA	50.14	49.15	80.37	78.23	0.683	0.594	–
MTL	–	–	80.50	78.80	–	–	–
TFN	49.80	–	79.40	79.70	0.610	0.671	Aligned
EF-LSTM	50.01	51.16	77.84/80.79	78.34/80.67	0.601	0.682	Aligned
MFN	51.34	52.76	78.94/82.86	79.55/82.85	0.573	0.718	Aligned
DFG	51.37	52.69	81.28/83.48	81.48/**83.23**	0.575	0.713	Aligned
MARN	–	–	–	–	–	–	–
MCTN	49.60	–	79.80	80.60	0.609	0.670	Aligned
RAVEN	50.00	–	79.10	79.50	0.614	0.662	Aligned
FmlMSN-m$_1$	**51.64**	**53.02**	**83.67/83.71**	**83.39**/83.17	**0.565**	**0.728**	Aligned
LF-DNN	50.83	51.97	80.60/82.74	80.85/82.52	0.580	0.709	Unaligned
TFN	51.60	53.10	78.50/81.89	78.96/81.74	0.573	0.714	Unaligned
LMF	51.59	52.99	80.54/83.48	80.94/83.36	0.576	0.717	Unaligned
CTC+MCTN	48.20	–	79.30	79.70	0.631	0.645	Unaligned
CTC+RAVEN	45.50	–	75.40	75.70	0.664	0.599	Unaligned
MulT	52.84	54.18	81.15/84.63	81.56/84.52	0.559	0.733	Unaligned
MISA	52.05	53.63	80.67/84.67	81.12/84.66	0.558	0.752	Unaligned
Self-MM	**53.87**	**55.53**	**83.76/85.15**	**83.82/84.90**	**0.531**	**0.765**	Unaligned
FmlMSN-m$_2$	52.69	54.17	83.45/84.29	83.56/84.10	0.569	0.719	Unaligned

Table 12
Some example videos for sentiment analysis. The second to third columns are the results obtained from three stages. Each has both bimodal task and unimodal tasks. M represents multimodal prediction. The right of '/' means wrong predicted label, while the left of that is the true label.

Utterances	$A+V$	A	V	$A+T$	A	T	$V+T$	V	T	M
这事结婚前咱俩不是说好了 Before we got married, didn't we both agree this thing	Neg	Neg	Neg	Neg/Pos	Neg	Neg/Pos	Neg	Neg	Neg	Neg
你妈说来磊儿来了影响方一凡学习 Your mother said that Leier will influence Fang Yifan to study	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg
那也让我见识下你的本事 That also allows me to see your skills	Neg	Pos/Neg	Neg	Pos	Pos/Neg	Pos	Neg/Pos	Neg	Pos	Neg/Pos
以我对你的判断我觉得你应该会先派黄橙橙 In my judgment on you, I think you should send Huang Chengcheng first	Pos	Pos/Neg	Pos	Pos	Pos	Pos	Pos	Pos	Pos	Pos
镇长，您把心放宽，多保重 Mayor, relax your mind and take care of yourself	Pos/Neg	Pos/Neg	Pos/Neg	Pos	Pos/Neg	Pos	Pos	Pos/Neg	Pos	Pos
无所谓，乱世 It doesn't matter, chaos world	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg
老马，平时在里边儿都爱喝啥啊 Old horse, what do you like to drink in it	Pos	Pos/Neg	Pos	Pos	Pos/Neg	Neg/Pos	Pos	Pos	Neg	Pos
现在好了，汽水不冰了，可我的心却是冰冰的 Well now, soda is not frozen, but my heart is frozen	Neg	Neg	Neg	Neg/Pos	Neg	Neg/Pos	Neg	Neg	Neg/Pos	Neg
倒是也行，你是法人，你看着办 It's OK. You're a legal person. You can do it	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg	Neg

see that FmlMSN-m$_1$ achieves better results on the most metrics than other methods. However, compared with unaligned models, FmlMSN-m$_2$ outperforms LF-DNN, TFN, LMF, MCTN, and RAVEN, but it is not as competitive on some metrics as those of MulT, MISA, and Self-MM. The gap is even more obvious on the MOSI dataset due to the small size of this dataset. Self-MM gets better results as it tends to automatically generate unimodal labels for multi-task learning, and applies the new strategy to balance different tasks. Compared with MulT and MISA, our model does not attend to optimize modality features representation. It is a kind of task-driven model. In this sense, our model still requires improvements like applying the attention mechanism in the submodule.

5.7. Case study and error analysis

We perform the case study and error analysis on the predictions of our proposed FmlMSN model. Some examples are listed in Table 12 and the results of each task are reported. Utterances mean the text in two different languages, of which the upper line is Chinese, and the lower line is the corresponding to English. A, V, and T represent the predicted labels of audio, video, and text, respectively. $A + V$, $A + T$, and $V + T$ represent the prediction of three bimodal tasks. M is multimodal prediction. The right of '×/×' means wrong predicted label, while the left of that is the true label. We mark the data in red font. Take the utterance 'Before we got married, didn't we both agree this thing' in Table 12 as an example, the 'Neg' indicates that the output of the proposed model and true label both are negative, the 'Neg/Pos' indicates that the prediction of AT has errors, where the predicted label is negative, but the true label is positive. For the unimodal tasks, it can be seen that the prediction of A has more errors. This is due to the noise in audio such as background music. The errors caused by T are because of the sarcasm and implicit sentiment. In these cases, V seems to achieve better performance. For the bimodal tasks, the performance of AT is not so satisfying. It is worth noting that even though some unimodal tasks have predicted wrongly, it does not affect the prediction of bimodal task and multimodal task. Take 'Old horse, what do you like to drink in it' as an example, A, T don't predict correctly in the stages of AV and AT, but the results of $A+V$ and $A+T$ and M are still right. The reason for these results is that the FmlMSN regards modalities of different granularities as independent tasks for multi-stage training based on fine-grained modal labels.

6. Conclusion

In consideration that each modality in the video expresses is heterogeneous and may express different sentiment, a fine-grained modal label-based multi-stage network is proposed for multimodal sentiment analysis. It exploits the sentiment labels in unimodal, bimodal, and multimodal to explore the information at different modal granularities and improve the generalization of the model. Besides, the irregular videos in CH-SIMS are reprocessed and the new features are provided for other researchers to use. The extensive experimental results show that our model outperforms the existing state-of-the-art methods. In future, we will study the strategies to balance different tasks and explore the importance of each stage. Besides, we will further mine the potential of ensemble learning in Multimodal sentiment analysis.

CRediT authorship contribution statement

Junjie Peng: Conceptualization, Writing – review & editing, Supervision. **Ting Wu:** Conceptualization, Methodology, Software, Writing – original draft. **Wenqiang Zhang:** Conceptualization, Writing – review & editing. **Feng Cheng:** Conceptualization, Resources. **Shuhua Tan:** Conceptualization, Resources. **Fen Yi:** Conceptualization, Resources. **Yansong Huang:** Formal analysis, Visualization.

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 data in the manuscript.

Acknowledgments

The authors would like to thank the funding from the Open Project Program of Shanghai Key Laboratory of Data Science (No. 2020090600004) and the resources and technical support from the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600).

References

Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion*, *76*, 204–226.

Akhtar, M. S., Chauhan, D. S., & Ekbal, A. (2020). A deep multi-task contextual attention framework for multi-modal affect analysis. *ACM Transactions on Knowledge Discovery Data*, *14*(3), 1–27.

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for Multi-modal emotion recognition and sentiment analysis. In *Proc. NAACL HLT - conf. N. AM. chapter assoc. comput. linguistics: hum. lang. technol.* (pp. 370–379).

Amir, Z., Paul, P. L., Soujanya, P., Prateek, V., Erik, C., & Louis-Philippe, M. (2018). Multi-attention recurrent network for human communication comprehension. In *The thirty-second AAAI conference on artificial intelligence* (pp. 5642–5649).

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources Evaluation*, *42*(4), 335–359.

Cao, R., Ye, C., & Hui, Z. (2021). Multimodel sentiment analysis with self-attention. In *FTC - proc. future technol. conf.* (pp. 16–26).

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for Multi-modal sentiment and emotion analysis. In *Proc. EMNLP - conf. empir. methods nat. lang. process conf.* (pp. 5646–5656).

Chauhan, D. S., Dhanush, S. R., Ekbal, A., & Bhattacharyya, P. (2020a). All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proc. IJCNLP - int. jt. conf. nat. lang. process.* (pp. 281–290).

Chauhan, D. S., Dhanush, S. R., Ekbal, A., & Bhattacharyya, P. (2020b). Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 4351–4360).

Chen, G., Peng, J., Zhang, W., Huang, K., Cheng, F., Yuan, H., & Huang, Y. (2021). A region group adaptive attention model for subtle expression recognition. *IEEE Transactions on Affective Computing*, 1–14. http://dx.doi.org/10.1109/TAFFC.2021.3133429.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL HLT - conf. N. Am. chapter assoc. comput. linguistics: hum. lang. technol.* (pp. 4171–4186).

Fortin, M., & Chaib-draa, B. (2019). Multimodal sentiment analysis: A multitask learning approach. In *Proc. lect. notes comput. sci.* (pp. 368–376).

Gaye, B., Zhang, D., & Wulamu, A. (2021). A tweet sentiment classification approach using a hybrid stacked ensemble technique. *Information*, *12*(9), 374.

Ghosal, D., Akhtar, M. S., Chauhan, D. S., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In *Proc. EMNLP - conf. empir. methods nat. lang. process conf.* (pp. 3454–3466).

Graves, A., Fernández, S., Gomez, F. J., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML - int. conf. mach. learn. conf.* (pp. 369–376).

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., & Marsic, I. (2018). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 2225–2235).

Han, W., Chen, H., Gelbukh, A. F., Zadeh, A., Morency, L., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. CoRR, abs/2107.13669.

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proc. MM - proc. ACM int. conf. multimed.* (pp. 1122–1131).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *International Journal of the Multimedia Information Retrieval*, *9*(2), 103–112.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. (2014). Large-scale video classification with convolutional neural networks. In *Proc. IEEE comput soc conf comput vision pattern recognit conf.* (pp. 1725–1732).

Kaur, R., & Kautish, S. (2019). Multimodal sentiment analysis: A survey and comparison. *International Journal of the Service Science Management and Engineering Technology*, *10*(2), 38–58.

Kazmaier, J., & van Vuuren, J. H. (2022). The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*, *187*, Article 115819.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In I. Gurevych, & Y. Miyao (Eds.), *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 2247–2256).

Mai, S., Hu, H., & Xing, S. (2020). Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proc. AAAI - artif. intell. conf.* (pp. 164–172).

Mai, S., Hu, H., Xu, J., & Xing, S. (2022). Multi-fusion residual memory network for multimodal human sentiment comprehension. *IEEE Transactions on Affecting Computers*, *13*(1), 320–334.

Mai, S., Xing, S., & Hu, H. (2021). Analyzing multimodal sentiment via acoustic- and visual-LSTM with channel-aware temporal convolution network. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1424–1437.

Majumder, N., Hazarika, D., Gelbukh, A. F., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, *161*, 124–133.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *8*, In *Proc. python in science conference* (pp. 18–25).

Pham, H., Liang, P. P., Manzini, T., Morency, L., & Póczos, B. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proc. AAAI - AAAI conf. artif. intell. conf.* (pp. 6892–6899).

Poria, S., Cambria, E., & Gelbukh, A. F. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proc. EMNLP - conf. empir. methods nat. lang. process conf.* (pp. 2539–2544).

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. (2017a). Context-dependent sentiment analysis in user-generated videos. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 873–883).

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. (2017b). Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Proc. IEEE int. conf. data min. ICDM* (pp. 1033–1038).

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Proc. IEEE int. conf. data min. ICDM* (pp. 439–448).

Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, *261*, 217–230.

Sahu, S., Mitra, V., Seneviratne, N., & Espy-Wilson, C. Y. (2019). Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription. In *Proc. annu. conf. int. speech. commun. assoc., INTERSPEECH* (pp. 3302–3306).

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. In *Proc. adv. neural inf. proces. syst. conf.* (pp. 525–536).

Tang, J., Li, K., Jin, X., Cichocki, A., Zhao, Q., & Kong, W. (2021). CTFN: hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 5301–5311).

Tian, L., Lai, C., & Moore, J. D. (2018). Polarity and intensity: the two aspects of sentiment analysis. CoRR, abs/1807.01466.

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 6558–6569).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proc. adv. neural inf. proces. syst. conf.* (pp. 5998–6008).

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proc. AAAI - AAAI conf. artif. intell. conf.* (pp. 7216–7223).

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing emotions in video using multimodal dnn feature fusion. In *Proc. challenge-HML. conf.* (pp. 11–19).

Wu, L., Liu, Q., Zhang, D., Wang, J., Li, S., & Zhou, G. (2020). Multimodal emotion recognition with auxiliary sentiment information. *Beijing Da Xue Xue Bao*, *56*(1), 75–81.

Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., & Huang, Y. (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems*, Article 107676.

Xi, C., Lu, G., & Yan, J. (2020). Multimodal sentiment analysis based on multi-head attention mechanism. In *Proc. ACM int. conf. proc. ser. conf.* (pp. 34–39).

Xu, Q., Peng, J., Zheng, C., Tan, S., Yi, F., & Cheng, F. (2023). Short text classification of chinese with label information assisting. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 1–14. http://dx.doi.org/10.1145/3582301.

Yadollahi, A., Shahraki, A. G., & Za, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Survey*, *50*(2), 1–33.

Yang, B., Wu, L., Shao, B., Lin, X., & Liu, T.-Y. (2022). Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2015–2024.

Ye, X., Dai, H., Dong, L., & Wang, X. (2021). Multi-view ensemble learning method for microblog sentiment classification. *Expert Systems with Applications*, *166*, Article 113987.

Yu, J., Jiang, J., & Xia, R. (2020). Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, *28*, 429–439.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 3718–3727).

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. CoRR, abs/2102.04830.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proc. EMNLP - conf. empir. methods nat. lang. process conf.* (pp. 1103–1114).

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. (2018). Memory fusion network for multi-view sequential learning. In *Proc. AAAI - AAAI conf. artif. intell. conf.* (pp. 5634–5641).

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proc. ACL - annu. meet. assoc. comput. linguist. conf.* (pp. 2236–2246).

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligence Systems*, *31*(6), 82–88.

Zhang, Y., & Yang, Q. (2017). A survey on multi-task learning. CoRR, abs/1707.08114.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.