FMSA-SC: A Fine-Grained Multimodal Sentiment Analysis Dataset Based on Stock Comment Videos

Lingyun Song, Siyu Chen, Ziyang Meng, Mingxuan Sun, and Xuequn Shang

***Abstract*—Previous Sentiment Analysis (SA) studies have demonstrated that exploring sentiment cues from multiple synchronized modalities can effectively improve the SA results. Unfortunately, until now there is no publicly available dataset for multimodal SA of the stock market. Existing datasets for stock market SA only provide textual stock comments, which usually contain words with ambiguous sentiments or even sarcasm words expressing opposite sentiments of literal meaning. To address this issue, we introduce a Fine-grained Multimodal Sentiment Analysis dataset built upon 1,247 Stock Comment videos, called FMSA-SC. It provides both multimodal sentiment annotations for the videos and unimodal sentiment annotations for the textual, visual, and acoustic modalities of the videos. In addition, FMSA-SC also provides fine-grained annotations that align text at the phrase level with visual and acoustic modalities. Furthermore, we present a new fine-grained multimodal multi-task framework as the baseline for multimodal SA on the FMSA-SC.**

***Index Terms*—Multimedia databases, neural networks, sentiment analysis, video signal processing.**

I. Introduction

THE stock market acts as an important investment channel and is essential for the development of industries that have a major impact on the economy. The perception of stock price risk is an important guarantee for the stable development of stock market. Existing works [1], [2] have demonstrated that Sentiment Analysis (SA) plays an important role in stock market prediction, e.g., stock price movement [3], [4] and volatility [5], [6].

In recent years, media podiums and commercial financial websites, such as *TikTok* and *Dianzhang*,[1] provide platforms for investors to share their opinions and reviews on recent trend of the stock market through online video postings. Analyzing the sentiments expressed in these videos helps identify the sentiments (positive and negative) of the stock market, which is a key factor in perceiving the movement of the stock market [7], [8].

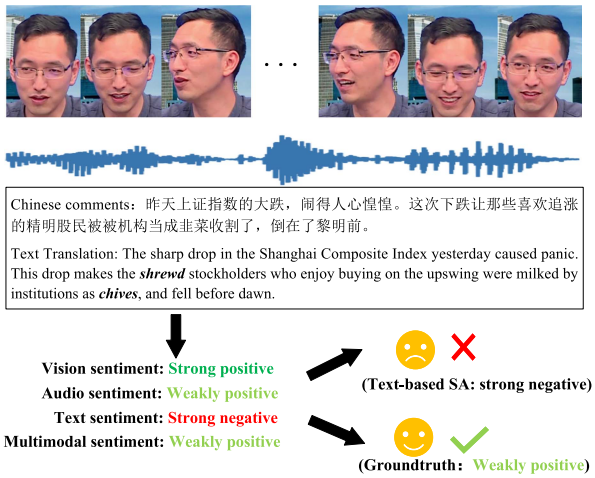


Fig. 1. Example of the sentiment difference between different modalities in a comment video for the stock market. Note that all sentiment labels of different modalities are manually annotated by workers. Textual comments contain ironic words such as “shrewd” and “chives”. The person in this video speaks Chinese Mandarin and the textual comments are translated into English for convenient reading.

Analyzing the sentiments of these stock comment videos is non-trivial. Most existing SA methods for the stock market only mine sentiment polarity from the contents of textual comments, overlooking the sentiment clues of visual and acoustic modalities. For example, some works [9], [10] analyzed the sentient polarities of the stock market based on text messages posted on social media (e.g., *Twitter*) or stock investment communities (e.g., *Stocktwits*). These methods overlook the exploration of sentiment cues in visual and acoustic modalities, which, if ignored, can result in false sentiment analysis results on videos. Taking the stock comments in Fig. 1 as an example, the textual comments contain the words indicating negative sentiments, such as “sharp drop” and “panic”. However, the smile and excited voice suggest positive sentiments of the speakers, which if overlooked, would lead to incorrect SA results. Hence, taking into account the sentiment information from all modalities in a comprehensive manner helps in accurately evaluating the sentiments expressed in videos.

Manuscript received 30 July 2023; revised 13 November 2023 and 27 January 2024; accepted 31 January 2024. Date of publication 8 February 2024; date of current version 24 April 2024. This work was supported in part by the National Nature Science Foundation of China under Grant 62102321, in part by the National Key Research and Development Program of China under Grant 2020AAA0108504, and in part by the Fundamental Research Funds for the Central Universities under Grant D5000230095. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guoying Zhao. *(Corresponding authors: Lingyun Song; Xuequn Shang; Ziyang Meng.)*

Lingyun Song, Siyu Chen, Ziyang Meng, and Xuequn Shang are with the School of Computer Science, Northwestern Polytechnical University, Xi’an 710129, China, and also with the Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi’an 710129, China (e-mail: lysong@nwpu.edu.cn; sychen@mail.nwpu.edu.cn; mzy@mail.nwpu.edu.cn; shang@nwpu.edu.cn).

Mingxuan Sun is with the Division of Computer Science and Engineering, School of Electrical Engineering and Computer Science, Louisiana State University, 70803 USA (e-mail: msun@csc.lsu.edu).

Data and codes are available at https://github.com/sunlitsong/FMSA-SC-dataset.git.

Digital Object Identifier 10.1109/TMM.2024.3363641

[1][Online]. Available: https://www.aniu.tv/.

Previous studies [11], [12], [13] have demonstrated that visual or acoustic data can provide useful sentiment cues to resolve the ambiguity and misunderstanding issues caused by text-based SA [14]. For example, the same word can vary in sentiment when paired with different acoustic or visual cues. Therefore, one promising solution to improve the precision of sentiment polarity analysis in videos is to combine sentiment cues from multiple modalities of videos [15], [16].

Although some efforts [14], [17], [18], [19] have been devoted to exploring multimodal sentiment cues, most of them focus on analyzing the sentiments of product reviews and public opinions, with little attention being paid to the sentiment analysis of stock market investors. Although promising performance is achieved, existing multimodal SA models are mainly built and trained on video blogs posted on social media, such as YouTube. Videos in different domains usually have differences in the characteristics of feature space and data distribution, which prevents the application of SA models trained with videos of other domains on the SA task of the stock market. For example, the sentiment expression characteristics of speakers talking about a product or a movie in these videos are different from those of stock market investors. As shown in Fig. 1, when talking about the stock market, investment experts often use figurative or ironic words with modest facial expressions. By contrast, consumers or talk show hosts usually have a talk with exaggerated expression or sarcastic laughter. This makes the multimodal SA models trained with videos from other fields unable to adapt well to stock market comment videos. In recent years, existing works [20], [21] also have demonstrated that when the data distribution/characteristics of the target domain are significantly different from those of the pre-training dataset from other domains, pre-trained large models (e.g., CLIP [22], [23]) cannot achieve satisfactory performance if they are not fine-tuned with the data of the target domain. Therefore, it is necessary to construct MSA datasets for the stock market.

Unfortunately, to our knowledge, there are no public multimodal Sentiment Analysis (MSA) datasets for the stock market. Most of the existing MSA datasets are mainly constructed based on the videos collected from movies, TV series, or video blogs posted on video sharing websites, such as the MOSEI [24] and MOSI [15] datasets. This hinders the establishment of specialized SA models for stock comment videos, which limits the accuracy of SA-based investment decision-making [25], [26] and risk perception tasks such as stock price volatility [27], [28].

To solve the above issue, we collect online videos talking about the stock market of China from a popular financial media platform *DianZhang Finance*[2] and establish a new fine-grained MSA dataset based on these videos, namely FMSA-SC. Unlike most of the existing MSA datasets, where only one unified multimodal sentiment label is annotated for each video, our FMSA-SC contains both multimodal and independent unimodal sentiment annotations for each video. This allows for studying unimodal SA tasks, the interactions between modalities for the MSA, and the multi-task learning based on both unimodal and multimodal SA tasks. Specifically, FMSA-SC has 1,247 stock comment videos collected from public media platforms and each video is used to generate synchronized three types of modality data, including text, audio and vision data. In summary, FMSA-SC has the following two characteristics:

1) To our knowledge, FMSA-SC is the first MSA dataset for the stock market, which provides unimodal sentiment annotations for three independent modalities (visual, acoustic and textual) and one unified multimodal annotation for each video. The fine-grained unimodal sentiment annotations are beneficial for learning to resolve the sentiment ambiguity brought by text-based SA. By contrast, existing SA datasets [9], [10] for the stock market only contain text comments, whereas MSA datasets in the fields being irrelevant to stocks usually only have one unified multimodal sentiment label, such as MuSe-CaR [29] and MOSEI [24].
2) FMSA-SC provides fine-grained alignments among different modalities at the phrase level. Specifically, we further segment videos based on the word segmentation marks (i.e., time stamps) of the texts transcribed from the videos. As a result, we obtain synchronized text, audio, and acoustic segments, as well as phrase-level multimodal alignments. Benefiting from these fine-grained alignment annotations, researchers are allowed to study new stock market SA methods that explore and aggregate sentiment cues from phrase-level data of different modalities. The entire data set will be available to the research community.

Based on FMSA-SC, we propose a **F**ine-**G**rained **M**ultimodal Multi-task framework as the baseline for stock market **S**entiment **A**nalysis, namely FGMSA. It addresses the MSA for the stock market by combining unimodal and multimodal SA tasks in a unified neural network, which facilitates the learning of modality representations that capture sentiment cues from different modalities. When learning multimodal sentiments, FGMSA assigns the phrase-level information of each modality a learnable importance weight based on interactions among different modalities. This allows each modality aligned to different phrases to flexibly exert different effects on multimodal sentiments. The motivation behind this is that each modality of a video usually conveys complex information, and not all the information is necessarily representative for sentiments. Besides, the sentiments of one modality can vary with the information of other modalities synchronized with it. For example, when the word "great" is paired with laughter in different tones, it can indicate either a positive sentiment like praise or a negative sentiment like sarcasm. Unimodal and multimodal tasks can enhance each other by jointly optimizing the learning of phrase-level features that capture sentiment cues from different modalities.

Extensive experiments and ablations on FMSA-SC show the advantages of FGMSA over previous state-of-the-art methods on the MSA task of the stock market.

II. Related Work

A. MSA Datasets

In recent years, researchers have proposed some multimodal datasets for the SA task, where multiple modalities are usually

[2][Online]. Available: https://www.aniu.tv/.

extracted from videos or multimodal messages posted on social media platforms. For example, an early well-known MSA dataset is the *YouTube Opinion Dataset* [30], which contains 47 YouTube videos with sentiment polarity annotations. The experiments based on this dataset verified the effectiveness of jointly exploring sentiment cues from different modalities for the MSA task. Soujanya et al. [31] proposed a MSA dataset for multi-party dialogues, where 1,433 dialogues encompassing acoustic, visual, and textual modalities are collected from the TV-series *Friends*. Pérez et al. [32] established a Spanish MSA dataset, where 105 YouTube videos are segmented into utterance-level video segments with sentiment polarity. Zadeh et al. [15] proposed a multimodal opinion-level sentiment intensity dataset, in which 93 YouTube videos are segmented into 2,199 opinion-level segments with sentiment annotations. Morency et al. [33] established a large MSA dataset consisting of 3,228 monologue videos collected from YouTube, where the 3 most frequent topics of these videos are reviews, debate and consulting. Lucia et al. [34] constructed a new MSA dataset by tweets containing texts and images, where each tweet is assigned one of the sentiment polarities, including positive, neural, and negative. Chauhan et al. [35] proposed a MSA dataset for multi-modal conversational context that provides sentiment annotations for acoustic-visual utterances of TV video clips.

The aforementioned MSA datasets contain sentiment attitudes towards diverse topics in various domains, including politic opinions, movie and product reviews, etc. However, few datasets focus on the sentiments of the stock market. By contrast, our FMSA-SC dataset is established by the stock comment videos, which is specifically designed to detect the sentiment polarities of the stock market. Besides, most of existing MSA datasets only have overall multimodal sentiment annotations for video clips, whereas FMSA-SC has both mulitmodal and independent unimodal sentiment annotations, as well as phrase-level alignments between different modalities. Although the recent MSA dataset CH-SIMS [12] also has both multimodal and unimodal sentiment annotations, it only contains 60 videos in the wild, which is irrelevant to stocks and does not contain cross-modal alignments at the phrase level.

B. MSA Methods

In many scenarios, text alone is not sufficient to accurately detect sentiment polarity. Existing MSA methods [36], [37], [38], [39] usually improve the accuracy of SA results by exploring sentiment cues from visual and acoustic modalities to complement textual sentiment cues. In these methods, various multimodal fusion methods are developed to fuse sentiment cues from multiple different modalities, such as feature concatenation [40], tensor factorization [41], [42] or importing attention mechanisms [36], [43], [44]. For example, Zadeh et al. [41] proposed a tensor fusion network that aggregates unimodal, bimodal and trimodal interactions for MSA. Wang et al. [37] proposed a cross-modal enhancement network that facilitates the integration of sentiment cues of different modalities by a feature transformation strategy. Liu et al. [42] proposed a low-rank tensor-based fusion method for MSA. Ji et al. [39] proposed a bilayer multimodal hypergraph learning to analyze the sentiments of multimodal tweets, where sentiment cues from different modalities are aggregated by feature-level hypergraph learning. Guo et al. [38] propose a multimodal attention network driven by layout to recognize sentiments in news, where the layout of online news is exploited to align images with the corresponding text to learn affective representations. Ashima et al. [45] proposed a deep multi-level attentive network for MSA, which learns multimodal features that capture rich sentiment cues from image and text modalities by a self-attention mechanism. Wang et al. [46] proposed learning multimodal features based on tripartite interactions for MSA by a nonuniform attention network. Zhu et al. [11] proposed a new image-text interaction network for MSA, which integrates sentiment information from image regions and textual words via an adaptive cross-modal gating module.

Although improving the SA results, these studies do not distinguish the influences of the same modality on the SA in different time periods. For example, the influences of language modality may vary with the utterance contexts at different time periods, exerting positive or negative influences on the SA results. By contrast, our proposed FGMSA assigns learnable weights to each modality aligned with different phrases, which can flexibly control the effects of each modality on fine-grained phrase-level sentiment learning.

Besides, these works only use unified mulitmodal sentiment annotations, which have limitations in capturing differentiated sentiment cues from different modalities. To address this issue, [47] proposed a self-supervised multi-task learning strategy for jointly training the multimodal and unimodal SA tasks, which generates pseudo unimodal sentiment supervisions to improve the learning of unimodal representations that contain modality-specific and complementary information. [12] established a MSA dataset with both unimodal and multimodal sentiment annotations and proposed a multi-task learning framework to aid the learning of complementary unimodal representations. Inspired by these methods, we propose a baseline for our MSA dataset (i.e., FMSA-SC) that integrates multimodal and unimodal SA tasks in an unified framework. Different from previous multi-task frameworks that directly explore global multimodal sentiment, our method explores the fine-grained multimodal sentiment by integrating sentiment cues of different modalities at the phrase level. This benefits from our FMSA-SC that provides phrase-level alignments between modalities.

III. FMSA-SC Dataset

A. Data Collection

Videos are collected from an online financial community and financial platform *Dianzhang Finance*,[3] where more than 300 well-known professional investors in finance and economics post videos. Videos from *Dianzhang* are recorded in MP4 format with high-tech microphones and cameras. The length of these stock comment videos varies from 1 to 2 minutes. In this work,

[3]https://www.aniu.tv/.

we first collect more than 10,000 raw stock review videos from the *Dianzhang* platform discussing the China stock market, from May 1, 2021 to December 25, 2021.

Then, we crop the raw videos into utterance-level video segments according to long pauses with an average of 10 utterances per video. After that, we collect high-quality video segments by the following filtering rules: 1) we select the video segments containing only one speaker, and the face and voice of the speaker need to appear at the same time.

2) We select the video segments talking about the stock market, and remove the video segments discussing irrelevant topics, such as praising other speakers.
3) We select Mandarin videos and are cautious when selecting video segments with accents.

Besides, as commentary utterances usually contain both subjective and objective comments, where the subjective comments express speakers' attitudes and objective comments only state the facts and truth [15], [48]. To obtain a high-quality MSA dataset for the stock market, we follow the work in [15] to further filter out the video segments of objective comments that do not express sentiments. Specifically, subjective comments containing private opinions of speakers should meet the following constraints.

1) The video segments contain explicit mention of a private attitude of the speakers. For example, video segments with the expression "I am optimistic about A-shares."
2) The video segments relay the attitudes of someone other than the speakers. For example, "Mr. Zhao believes that A-shares will soar in the near future."
3) The video segments contain no direct mentions of private attitude, but give information implying the attitudes. For example, "The A-share market has been booming recently."

Finally, we collect 1,247 video segments of the stock market and each segment is transcribed to extract three modalities, including text, audio and silent video clips. Text is aligned at the phrase level with audio and video clips. Specifically, our methodology for the transcription and the cross-modal alignment are performed by three stages: i) a speech recognition algorithm is first used to transcribe all the videos to texts, and then 3 experts review and correct all the transcribed texts. The texts contain the details about stresses and speech pause, but discard meaningless words such as "umm" and "uhh". ii) Texts are segmented by *Short Form ASR*[4] of the iFLYTEK Open Platform to obtain phrase-level segments and their timestamps. The segmentation results are reviewed by 3 experts to ensure their accuracy. iii) The transcribed texts are carefully aligned with the video clips and audio at the phrase level by the timestamp. The detailed statistics and word cloud map of our dataset are shown in Table I and Fig. 2, respectively.

TABLE I
STATISTICS OF OUR FMSA-SC DATASET

Item	#
Total number of video segments	1,247
Average duration of per segment	6.3s
Average sentence count per segment	2.6
Total number of phrases	26637
Total number of unique phrases	2832

Fig. 2. Word cloud of our FMSA-SC. The larger the size of the words, the more frequently the words appear.

B. Annotation

Our FMSA-SC dataset contains one unified multimodal sentiment label for each video segment and independent sentiment labels for all three modalities, i.e., textual, acoustic and visual modalities extracted from the video segments. To reduce the mutual interference between different modalities, the workers are asked to follow the rules below to ensure the high quality of the annotations.

- When annotating unimodal sentiment, each worker can only see one type of modality data and is asked to keep in mind that information from the previous modality is not allowed to annotate the next modality.
- Each worker makes unimodal annotations before multimodal annotations. Specifically, workers annotate textual data (i.e., texts) first, acoustic data (i.e., audios) second, then visual data (i.e., silent videos), and multimodal data (videos) last.

Specifically, for each video segment, we employ five workers from the professional data annotation platform *JingLianWen Technology*[5] to give unimodal and multimodal annotations. The selection criteria for these workers are that they have experience in stock investment, and have been engaged in data annotation for more than 3 years. Each worker has three choices of sentiment state when performing the annotation, including 1 (positive), 0 (neutral) and -1 (negative). For each sample, if the annotation results given by five workers show a large disagreement, e.g., $\sum(-1) \geq 2 \& \sum(1) \geq 2$, we ask another five workers to annotate the sample again. By averaging the annotation results given by five workers, the final annotation 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 annotations can be further divided into 5 categories for classification tasks: strong negative $\{-1.0, -0.8\}$, weakly

[4][Online]. Available: https://global.xfyun.cn/products/lfasr/.

[5][Online]. Available: http://www.jinglianwen.com/

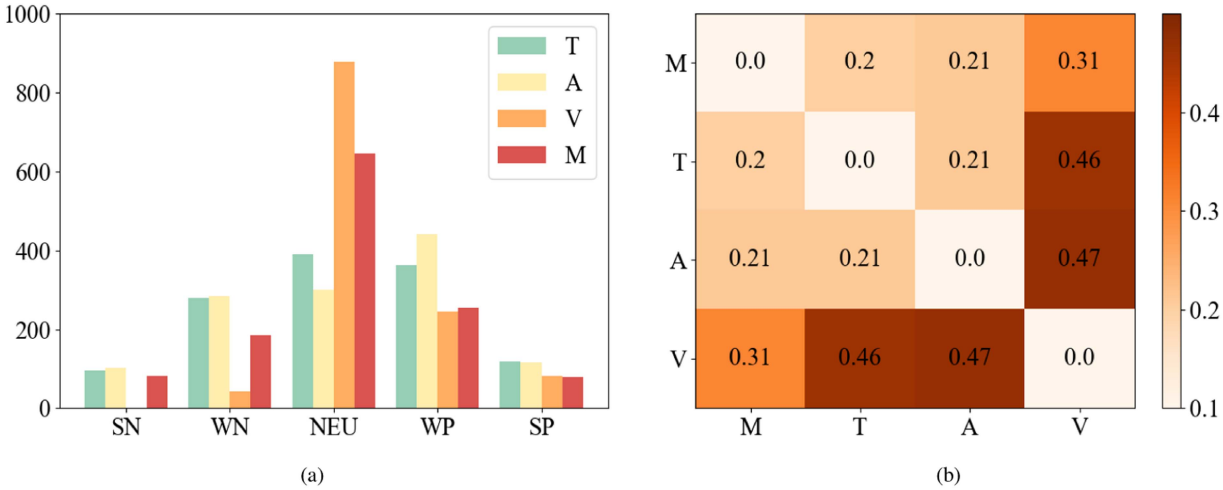


Fig. 3. (a) Sentiment distribution of **T**ext, **A**udio, **V**ideo and **M**ultimodality. (b) The sentiment confusion matrix indicates the differences in the sentiment annotations of different modalities in our FMSA-SC. The higher the value, the bigger the difference.

negative $\{-0.6, -0.4\}$, neutral $\{-0.2, 0.0, 0.2\}$, weakly positive $\{0.4, 0.6\}$, and strong positive $\{0.8, 1.0\}$. The agreement between annotation workers is 0.79 in terms of Krippendorff's Alpha [49], which is a measure of inter-annotator agreement that takes into account chance agreement.

To illustrate the difference in sentiment annotations of different modality data, in Fig. 3(b) we present a sentiment confusion matrix, which can be computed by

$$D_{ij} = \frac{1}{N}\sum_{n=1}^{N}\left(y_i^n - y_j^n\right)^2, \tag{1}$$

where $i, j \in \{t, a, v, m\}$ and t, a, v, m denote the Textual, Acoustic, Visual and Multimodal data, respectively. N denotes the number of video segments, and y_i^n denotes the n-th label for the i-th modality.

As seen in Fig. 3(b), compared with modality v, both modalities t and a have smaller difference with multimodal m. As multimodal annotations are deemed to be the sentiment truth, we can draw the following two insights. First, for stock comment videos, textual and acoustic modalities contain more valuable sentiment cues than the visual modality in some cases. Second, although the difference between t or a and m is small, only the textual or acoustic modality is still insufficient to guarantee accurate sentiment predictions. Combining the verbal cues in texts with the cues in vocal and visual expressions helps to accurately infer the sentiments of stock comment videos.

The difference between t and a is smaller than that between t and v. This suggests that the sentiments expressed in utterance and audio are closer, and they do not coincide with the sentiments of facial expressions in some cases. This is in line with expectation because audio contains text information but has sparse connections with vision information. In Fig. 3(a), we show the histogram of the sentiment distributions of unimodal and multimodal samples in our dataset. We have the following two observations:

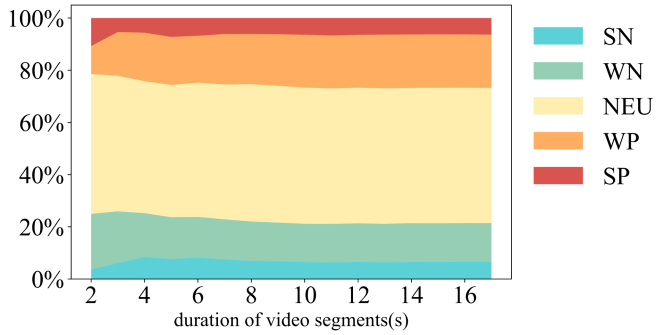


Fig. 4. Percentage of each sentiment category with different sizes of video segments. SN: Strong Negative, WN: Weakly Negative, NEU: Neutral, WP: Weakly Positive, SP: Strong Positive.

1) There is no significant difference in the number of multimodal samples that express positive and negative sentiments.
2) After introducing the visual modality into the SA of videos, the sentiment polarity of some videos is corrected. For example, compared to visual sentiments, acoustic and textual sentiments suggest that more videos are *WP*. However, the number of *WP* videos are reduced based on multimodal annotations, which introduce visual information into SA in addition to textual and acoustic information.

Fig. 4 shows the multi-modal sentiment distribution of all video segments whose duration is less than the value labeled on the x-axis. The maximum duration of the video segments in our dataset is 17 seconds. We can observe that the proportions are mostly consistent across all segment sizes. It indirectly proves that our data annotation is reliable.

C. Feature Extraction

In this part, we introduce the feature learning of each modality in our experiments.

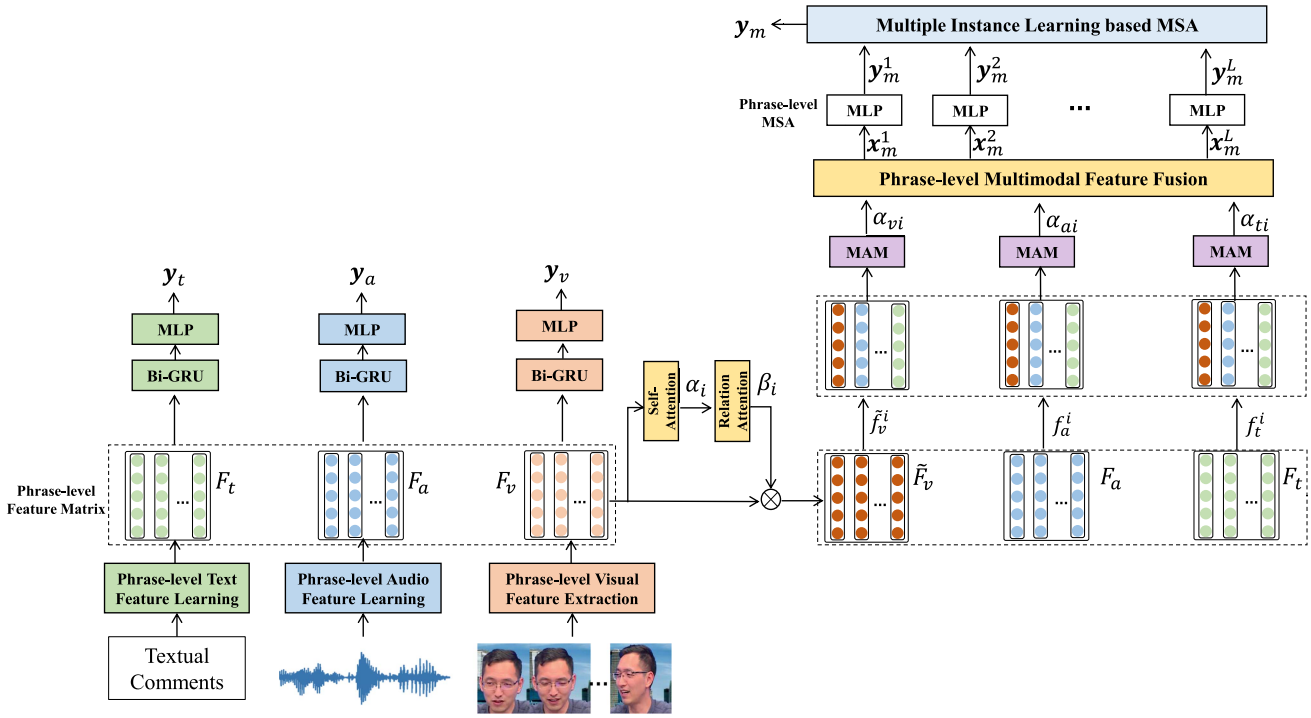


Fig. 5. Architecture of our FGMSA.

Textual feature: We learn textual features for the transcriptions of videos by a pre-trained BERT model[6] [50]. Specifically, we first learn a 768 dimensional vector for each word of the transcriptions by the BERT model. Then, we learn phrase-level textual representations by averaging the word vectors of all the words in each phrase.

Acoustic feature: We learn acoustic features for the audio track of videos at 16000 Hz by a Wav2vec 2.0 [51] model, which has been pre-trained on 10 k hours Speech data from Wenet-Speech [52]. As a result, we learn a 768 dimensional acoustic feature for each sampling point of the audio. Phrase-level acoustic features are learned by averaging the acoustic features of all sampling points of the audio aligned with each phrase.

Visual feature: We learn visual features for each frame of videos at 25 Hz by the Openface toolkit 2.0 [53]. Specifically, for each frame, we extract 136 facial landmarks, 35 facial action units, 288 dimensional eye gaze feature and 6 head poses. Finally, we obtain a set of 465-dimensional frame-level visual features. Phrase-level visual features are learned by averaging the visual features of all the frames aligned with the phrase.

Please note that as each phrase contains only a few words, the information loss caused by the average operation used in the learning of phrase-level modality features is small. During the small time interval, the expressions or acoustic characteristics usually do not have large variations in the corresponding frames or audio clips. Thus, we follow existing MSA works (e.g., baselines MAG-BERT [54] and LMF [42]) to use the average operation to obtain the phrase-level textual, visual and acoustic features.

[6][Online]. Available: https://huggingface.co/bert-base-chinese.

IV. Methodology

Fig. 5 shows the framework of the proposed fine-grained multimodal multi-task SA framework for the stock market, namely FGMSA. It integrates unimodal and multimodal SA tasks in an unified framework, where phrase-level features of different modalities are used for the corresponding unimodal SA and fused for MSA.

A. Unimodal Sentiment Prediction

In this part, we predict unimodal sentiment for three modalities, i.e., textual, acoustic and visual data, which can be formulated by

$$\boldsymbol{x}_u = \phi_u\left(\mathcal{S}_u(F_u)\right), \tag{2}$$

$$\boldsymbol{y}_u = \sigma(\boldsymbol{x}_u), \tag{3}$$

where $F_u \in \mathbb{R}^{L_u \times D_u}$ denotes the feature matrix for the sequence of modality $u \in \{t, a, v\}$, L_u denotes the phrase-level sequence length of modality u and D_u denotes the dimension of phrase-level features learned according to Section III-C. $\mathcal{S}_u(\cdot)$ denotes a sequence feature extractor for the modality u. In our experiments, the Bi-GRU network [55] is used as $\mathcal{S}_u(\cdot)$ to learn the sequence features of each modality. Note that Bi-GRU has been widely used in baseline methods to model sequence data and thus is used for fair comparisons. In fact, Bi-GRU can be exchanged by other sequence modeling methods, e.g., Transformer. $\phi_u(\cdot)$ denotes a multilayer perceptron consisting of two hidden layers with $ReLu$ as activation functions. $\boldsymbol{x}_u \in \mathbb{R}^{D'_u}$ denotes the feature representation for modality u and D'_u denotes the dimension of the features learned by $\phi_u(\cdot)$. $\boldsymbol{y}_u$ represents the predicted sentiment distribution for the modality u and $\sigma(\cdot)$ denotes the $Softmax$ function.

B. Multimodal Sentiment Prediction

In this part, we predict multimodal sentiment by combining sentiment cues from all modalities. As each modality that aligns with different phrases of utterances may contain sentiment clues of different polarity, it is necessary to first infer fine-grained phrase-level multimodal sentiments.

Due to the difference between the visual and multimodal sentiments described in Section III-B, it is necessary to first filter out the noisy information in the visual modality before phrase-level multimodal fusion. This is because although the visual modality contains valuable sentiment cues, not all of the video frames are representative for sentiments, which may contain noisy information that misleads SA results. Therefore, it is necessary to learn the importance of the visual modality aligned with each phrase, which helps to filter out the noisy visual information.

Following the attention mechanism in [56], we first learn self-attention weights α_i for the i-th phrase-level visual modality (i.e., silent video clips) based on its visual features. Specifically, α_i can be learned by

$$\alpha_i = Sigmoid\left(\boldsymbol{w}_1 \cdot \boldsymbol{f}_v^i\right), \tag{4}$$

where $\boldsymbol{f}_v^i$ denotes the visual feature for the silent video clip aligned with the i-th phrase and can be obtained according to Section III-C. The $\boldsymbol{w}_1$ denotes a learnable parameter vector. However, the learned weights α_i are coarse, because it only applies a nonlinear mapping to individual $\boldsymbol{f}_v^i$, overlooking the relations between local visual content aligned with phrases and global video content.

To refine α_i, we first aggregate all phrase-level visual features $\boldsymbol{f}_v^i$ to learn a global video-level feature by

$$\boldsymbol{f}_v = \frac{\sum_{i=1}^{L} \alpha_i \boldsymbol{f}_v^i}{\sum_i^n \alpha_i}, \tag{5}$$

where L denotes the number of the phrases in transcribed texts. Then, taking the global video-level feature $\boldsymbol{f}_v$ as an anchor, we further learn relation attention β_i for each phrase-level visual feature by modeling the relationships between $\boldsymbol{f}_v^i$ and $\boldsymbol{f}_v$. Inspired by the relation network in [57], we model the relationships between $\boldsymbol{f}_v^i$ and $\boldsymbol{f}_v$ by feeding their concatenations into a FC layer and learn β_i by

$$\beta_i = Sigmoid\left(\boldsymbol{w}_2 \cdot \left[\boldsymbol{f}_v^i, \boldsymbol{f}_v\right]\right), \tag{6}$$

where $\boldsymbol{w}_2$ is the parameter vector of the FC layer and the operator $[\cdot]$ denotes the concatenation of two vectors.

Finally, we learn new phrase-level visual features $\widetilde{\boldsymbol{f}}_v^i$ by using the coarse self-attention weights α_i and relation weights β_i to filter out noisy visual information, which can be formulated by

$$\widetilde{\boldsymbol{f}}_v^i = Relu\left(\boldsymbol{w}_3 \cdot \alpha_i \beta_i \left[\boldsymbol{f}_v^i, \boldsymbol{f}_v\right]\right), \tag{7}$$

where $\boldsymbol{w}_3$ is a trainable parameter matrix and $\boldsymbol{f}_v$ is used to provide global visual context information, which helps to enhance the phrase-level SA.

After obtaining all the $\widetilde{\boldsymbol{f}}_v^i$, we learn the multimodal feature $\boldsymbol{x}_m^i$ corresponding to the i-th ($1 \le i \le L$) phrase by

$$\boldsymbol{x}_m^i = \mathcal{F}\left(\alpha_{ti}\boldsymbol{f}_t^i, \alpha_{ai}\boldsymbol{f}_a^i, \alpha_{vi}\widetilde{\boldsymbol{f}}_v^i\right), \tag{8}$$

where L denotes the length of the text transcribed from a video segment and $\widetilde{\boldsymbol{f}}_v^i, \boldsymbol{f}_a^i, \boldsymbol{f}_t^i$ denote the features for the visual, acoustic and textual modality aligned with the i-th phrase. $\mathcal{F}(\cdot)$ is a feature fusion network. In our experiments, we report results using three different feature fusion methods, including LMF [42], Later Fusion DNN (LF-DNN) [12] and TFN [41]. α_{ui} denotes the attention weight for each phrase-level modality, indicating the importance of each modality for learning multimodal sentiment.

Specifically, $\alpha_{ui}(u \in \{t, a, v\})$ can be computed by a Multimodal Attention Mechanism (MAM), where the attention of one modality is adjusted by the interactions between the modality and other two modalities. This is consistent with the fact that the sentiment of one modality (e.g., word "great") can vary with other modality information (e.g, sarcastic tones) that synchronizes with it.

Specifically, taking the textual modality as an example, MAM learns the attention α_{ti} for the i-th phrase-level text feature by

$$\alpha_{ti} = w_t \cdot \eta_{ti} + w_a \cdot \eta_{ai} + w_v \cdot \eta_{vi} + b, \tag{9}$$

where η_{ti}, η_{ai} and η_{vi} denote the weights of phrase-level textual, acoustic and visual modalities, respectively, w_t, w_a and w_v are learnable weight parameters, and b is a bias. η_{ti}, η_{ai} and η_{vi} can be learned by

$$\eta_{ti} = Relu\left(\boldsymbol{q}_{ti} \cdot \boldsymbol{k}_{ti}^{\top}\right), \tag{10}$$

$$\eta_{ai} = Relu\left(\boldsymbol{q}_{ai} \cdot \boldsymbol{k}_{ai}^{\top}\right), \tag{11}$$

$$\eta_{vi} = Relu\left(\boldsymbol{q}_{vi} \cdot \boldsymbol{k}_{vi}^{\top}\right), \tag{12}$$

where $\boldsymbol{q}_{ti} = \boldsymbol{k}_{ti} = W_0\boldsymbol{f}_t^i$, $\boldsymbol{q}_{ai} = \boldsymbol{k}_{ai} = W_1\boldsymbol{f}_a^i$ and $\boldsymbol{q}_{vi} = \boldsymbol{k}_{vi} = W_2\widetilde{\boldsymbol{f}}_v^i$. W_0, W_1 and W_2 are three learnable parameter matrices. The attention weights α_{ai} and α_{vi} can also be learned using the method described above. Finally, we learn multimodal sentiment at the phrase-level by $\boldsymbol{y}_m^i = MLP(\boldsymbol{x}_m^i)$, where $MLP(\cdot)$ denotes an MLP with two hidden layers.

C. Multiple Instance Learning

After obtaining the set of fine-grained multimodal sentiment predictions $Y = (\boldsymbol{y}_m^1, \boldsymbol{y}_m^2, \ldots, \boldsymbol{y}_m^L)$, we learn the final global multimodal sentiment predictions for each video segment under the framework of Multiple Instance Learning (MIL) [58]. Specifically, we view each video segment as a bag of phrase-level multimodal instances. The bag-level multimodal sentiment $\boldsymbol{y}_m$ for each video segment can be learned by

$$\boldsymbol{y}_m = \sum_{i=1}^{l} \gamma_i \cdot \boldsymbol{y}_m^i, \tag{13}$$

where γ_i represents the weight for the i-th instance-level prediction $\boldsymbol{y}_m^i$. Specifically, γ_i can be learned by feeding the sequence $(\boldsymbol{x}_m^1, \boldsymbol{x}_m^2, \ldots, \boldsymbol{x}_m^L)$ into a Bi-GRU network, which can be formulated by

$$\gamma_i = \text{Softmax}\left(MLP\left(\left[\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i}\right]\right)\right), \tag{14}$$

where $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$ denote the forward and backward hidden state vectors learned by the Bi-GRU network. The $MLP(\cdot)$ denotes a two-layer MLP network that uses the $tanh$ activation.

D. Optimization Objectives

The three unimodal and one multimodal sentiment prediction tasks are jointly optimized by minimizing the following objective loss

$$\min \frac{1}{N}\sum_{n=1}^{N}\sum_{i}\lambda_i L\left(\boldsymbol{y}_i^n, \hat{\boldsymbol{y}}_i^n\right)+\sum_{j}\delta_j\|W_j\|_2^2, \qquad (15)$$

where N denotes the number of training samples, $i \in \{t, v, a, m\}$, $j \in \{t, v, a\}$, and λ_i and δ_j denote the hyper-parameters used for balancing different tasks and the L_2 norm, respectively. W_j denotes the parameters shared by the unimodal task of modality j and the multimodal task. $\boldsymbol{y}_i^n$ and $\hat{\boldsymbol{y}}_i^n$ denote the predicted and groundtruth sentiments for the n-th sample of modality i. $L(y_i^n, \hat{y}_i^n)$ denotes a L_1 loss for model training.

V. Experiment

A. Experimental Setup

Dataset Splits: We randomly shuffle all the videos and then split the videos into training, validation and testing sets according to the ratio of $6:2:2$. The detailed results of the dataset split are shown in Table III.

Baseline Methods: As shown in Table IV, we compare the performance of our FGMSA with previous state-of-the-art MSA baselines, falling into two main groups: i) Single-task learning-based methods. ii) Multi-task learning based methods. Specifically, *Self-MM* [47], *MLMF* [12], *MTFN* [12] and *MLF-DNN* [12] are multi-task learning based methods and the rest baselines are single-task learning based methods. Specifically, we introduce these baseline methods in detail as follows.

Early Fusion LSTM (EF-LSTM) [40]: This method makes MSA predictions based on multimodal features learned by concatenating input-level initial features of three different modalities.

Later Fusion DNN (LF-DNN) [59]: It learns features of different modality data first and then concatenates these unimodal features before making MSA predictions.

Low-rank Multimodal Fusion (LMF) [42]: It learns multimodal features for MSA based on modality-specific and cross-modal interactions.

Tensor Fusion Network (TFN) [41]: It learns multimodal features for MSA by aggregating unimodal, bimodal and trimodal interactions across textual, visual and acoustic modalities.

Multimodal Adaptation Gate-BERT (MAG-BERT) [54]: It uses the Multimodal Adaptation Gate (MAG) to receive multimodal information when fine-tuning BERT, which is achieved by modifying the internal representation of BERT. The MAG can be viewed as an attachment to the BERT that leaks multimodal information into the BERT.

Modality-Invariant and -Specific MSA (MISA) [60]: It projects each modality onto two distinct subspaces, that is, modality-invariant and modality-specific subspaces. Specifically, the first subspace is used to reduce the differences of the representations between modalities, whereas the second subspace is used to capture unique features for each modality.

Self-supervised Multi-task MSA (Self-MM) [47]: It first uses a self-supervised learning strategy to acquire independent pseudo-unimodal sentiment supervisions. Then, the multimodal and unimodal sentiment analysis tasks are jointly optimized in a unified framework.

Multimodal Information Maximization (MMIM) [61]: It is a hierarchical MI maximization framework for MSA, where MI maximization is used at both the input level and the fusion level in a multimodal fusion pipeline.

Multimodal Transformer (MulT) [62]: It is an extended Transformer network designed to learn representations from unaligned multimodal streams. Specifically, with multiple directional pairwise cross-modal transformers, MulT can merge multimodal time-series features for MSA via a feed-forward fusion process.

Multimodal multi-task learning framework [12]: It is a popular multimodal and multi-task learning framework for MSA, where one multimodal and three unimodal sentiment analysis tasks are jointly optimized with corresponding sentiment supervisions. This work presents three new MSA methods, i.e., MTFN, MLMF and MLF-DNN, that learn intermodal representations via different late-fusion strategies under the multi-task framework.

Metric: Following the baseline methods [12], we evaluate the performance of different methods in two forms: multi-class classification and regression. For multi-class classification, we report the F1 score and multi-class accuracy Acc-k ($k \in \{2, 3, 5\}$). For regression, we report Mean Absolute Error (MAE), where smaller values indicate better performance and zero is considered perfect. Except for MAE, higher values denote better performance for all metrics.

Implementation Details: All of our experiments are conducted exclusively on the NVIDIA GPU A40 and the Intel(R) Xeon(R) Gold 6326 CPU.

All the implementations of baseline methods come from the open-source MSA framework M-SENA [63]. We use a unified train-val-test split (i.e., 6:2:2) for fair comparisons between our method and competing methods. In our dataset, we align the textual, acoustic and visual modalities by the timestamp annotations for segmented phrases. Our model includes three unimodal and one multimodal SA tasks. During the training phase, we choose the hyper-parameters for each SA task from a grid search. For the final loss, the hyper-parameters λ_i and δ_j denote the weight parameters for $L(\boldsymbol{y}_i^n, \hat{\boldsymbol{y}}_i^n)$ and $\|W_j\|_2^2$, respectively. Specifically, λ_i ranges from 0 to 1, and we select λ_i with the step size 0.2. The δ_j is chosen from one of [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 0]. From the table, we can see that the optimal λ_i and δ_j for different variant models are different. For all variant models, each unimodal SA task and the MSA task contribute to the learning of optimal prediction results. This indicates that it is beneficial to combine unimodal and multimodal SA tasks in one unified framework to improve SA results. This is because unimodal and multimodal tasks can enhance each other by jointly optimizing the learning

TABLE II
OPTIMAL HYPER-PARAMETERS FOR OUR VARIANT MODELS

Model	MLP_layer	Number of neurons				Loss						
		T	A	V	M	λ_t	λ_a	λ_v	λ_m	δ_t	δ_a	δ_v
FGMSA-TFN	L1_layer	64	32	32	16	1.0	0.8	0.4	0.8	0	0	0
	L2_layer	64	32	32	16							
	L3_layer	1	1	1	1							
FGMSA-LMF	L1_layer	8	16	8	-	0.8	0.8	0.8	0.4	1e-5	1e-4	1e-5
	L2_layer	8	16	8								
	L3_layer	1	1	1								
FGMSA-LF-DNN	L1_layer	64	16	16	16	0.8	0.8	1.0	1.0	1e-4	1e-5	1e-4
	L2_layer	64	16	16	16							
	L3_layer	1	1	1	1							

TABLE III
DATASET SPLITS IN FMSA-SC

Item	SN	WN	NE	WP	SP	Total
#Train	49	111	388	153	47	748
#Valid	16	37	129	51	16	249
#Test	17	37	129	51	16	250

SN: Strong Negative, WN: Weakly Negative, NE: Neutral, WP: Weakly Positive, SP: Strong Positive.

of phrase-level modality features, which capture sentiment cues from different modalities.

In Table II we report the layer sizes (i.e., number of neurons) of the MLP used in three variants of our FGMSA that use different feature fusion methods. In the table, we separately report the sizes of the MLP used for Textual (T), Acoustic (A), visual (V) and Multimodal (M) SA tasks. Note that the original LMF model makes multimodal sentiment predictions by the product of the multimodal features and a learnable parameter, which leads to its best performance. Thus, in our FGMSA-LMF, we follow this method and do not use MLP for performing the MSA task.

During the training phase, we use the Adam solver with $\beta 1 = 0.9$, $\beta 2 = 0.999$ to optimize the parameters of our three models. The batch sizes for FGMSA-LMF, FGMSA-TFN and FGMSA-LFDNN are set to 64, 128 and 32 respectively, while their learning rates are 5e-4, 2e-3 and 2e-3. For a fair comparison, we report the average performance results of five times for each method.

B. Experimental Results

1) Comparisons With Baseline Methods: In this part, we first compare FGMSA with previous state-of-the-art MSA baselines on our FMSA-SC dataset. Then, to further verify the performance of our FGMSA, we conducted comparison experiments on the public MSA dataset CH-SIMS [12], which is currently the only publicly available MSA dataset with unimodal and multimodal sentiment annotations. It collects 2,281 video segments from movies, TV serials, and variety shows, but does not contain the videos discussing the stock market and phrase-level cross-modal alignment annotations. We employed workers who gave annotations for our FMSA-SC to annotate the phrase-level alignment results for CH-SIMS. The annotation results will be made public.

Specifically, the results of the comparison between FGMSA and the baselines on our FMSA-SC dataset are reported in Table IV. We can observe that FGMSA outperforms previous best methods in all metrics by a large margin, increasing the *Acc-2*, *Acc-5*, and *F1-score* by 4.86%, 6.07% and 3.71%, respectively. Specifically, with the same multimodal feature fusion method (i.e., LFM, TFN, and LF-DNN), FGMSA also achieves better performance than the corresponding multi-task MSA baselines. For example, compared with *MTFN*, our *FGMSA-TFN* achieves a 7.69% *Acc-2* improvement. The superior performance of our model over these baselines is mainly due to two reasons: i) our model can effectively explore phrase-level multimodal sentiments, which helps to mine subtle sentiment cues contained in short video clips. ii) Our model can adaptively suppress noisy information from phrase-level textual, acoustic, and visual modalities based on their interactions. This helps to reduce the negative influence produced by the inconsistency between different modalities. For example, when positive words (e.g., great) are uttered with exaggerated laughter or expressions, FGMSA can adaptively reduce the attention weights for the words in multimodal fusion, amplifying the acoustic or visual modality information that can reflect true sentiments. In contrast, previous methods learn multimodal features in a crusty way by combining entire features of different modalities.

To demonstrate the necessity of constructing specific MSA datasets for the stock market. We test the performance of a pre-trained large model without fine-tuning on our FMSA-SC. Specifically, we first train a large model MISA (denoted by MISA†) on the largest MSA dataset, i.e., MOSEI [24], which contains 23,453 annotated video segments from 250 topics (e.g., reviews and marketing). Then, we select one popular benchmark MSA method (i.e., MISA) that can achieve state-of-the-art performance on the CMU-MOSEI dataset. Then, we test the

TABLE IV
RESULTS (%) FOR SENTIMENT ANALYSIS ON OUR FMSA-SC DATASET

	Model	Acc-2	Acc-3	Acc-5	F1	MAE	Params	GFLOPs	Inf (ms)
Coarse-grained fusion	EF-LSTM [40]	68.02	43.32	43.32	56.26	29.56	0.28	0.03	0.52
	MulT [62]	69.64	55.06	44.94	70.64	28.50	87.18	9.15	40.90
	MAG-BERT [54]	72.06	55.47	47.77	72.91	29.01	88.73	8.90	18.08
	MISA† [60]	57.14	-	-	42.31	46.94	136.29	8.78	30.36
	MISA [60]	71.66	55.87	48.18	72.61	28.83	136.29	8.78	30.36
	Self-MM* [47]	69.23	53.85	48.99	70.18	28.08	86.01	8.64	35.23
	MMIM [61]	70.04	50.61	45.75	69.01	28.33	86.08	8.63	49.08
	LMF [42]	68.02	44.13	43.72	60.88	29.52	0.27	0.02	1.87
	MLMF* [12]	67.21	46.15	44.13	60.52	29.39	1.41	0.04	1.97
	TFN [41]	68.42	43.72	43.72	61.12	29.58	35.63	0.08	2.59
	MTFN* [12]	68.02	48.18	43.72	60.48	29.57	140.96	0.18	4.01
	LF-DNN [59]	68.02	43.32	43.32	57.72	29.56	0.64	0.02	0.71
	MLF-DNN* [12]	70.04	50.20	46.96	70.14	28.89	0.70	0.02	1.39
Phrase-level alignment	FGMSA-LMF	75.30	57.09	53.85	74.77	27.88	1.33	16.70	3.05
	FGMSA-TFN	75.71	57.09	**55.06**	74.89	27.74	3.58	16.77	2.97
	FGMSA-LF-DNN	**76.92**	**58.30**	**55.06**	**76.62**	**27.41**	2.46	35.90	3.11

The baselines with * are multi-task models, extended from single-task models by introducing independent unimodal annotations. The † indicates a large model that is pre-trained only on other large-scale MSA datasets, but not fine-tuned using our FMSA-SC data. Note that the lower the MAE value, the better the performance. Inf denotes the inference time that one model takes to infer the sentiment result for a video.SN: Strong Negative, WN: Weakly Negative, NE: Neutral, WP: Weakly Positive, SP: Strong Positive.

TABLE V
RESULTS (%) FOR SENTIMENT ANALYSIS ON SIMS DATASET

	Model	Acc-5
Coarse-grained fusion	EF-LSTM [40]	21.02
	MulT [62]	35.34
	LMF [42]	35.14
	MLMF* [12]	37.33
	TFN [41]	38.38
	MTFN* [12]	37.20
	MFN [64]	39.47
	LF-DNN [59]	41.62
	MLF-DNN* [12]	38.03
	MAG-BERT [54]	40.50
	Self-MM* [47]	41.53
Phrase-level alignment	FGMSA-LMF	35.84
	FGMSA-TFN	**45.35**
	FGMSA-LF-DNN	42.04

The baselines with * are the multi-task models.

performance of MISA† on the testing set of our FMSA-SC and report the result in Table IV. Compared with MISA that was fine-tuned using the FMSA-SC dataset, the performance of pre-trained MISA† occurs a sharp drop on all metrics, i.e., 14.52% for Acc-2, 30.3% for F1-score and 18.11% for MAE. We do not compare the performance by Acc-3, Acc-5 or Acc-7, because MOSEI divides sentiments into 2 categories or 7 categories and most of the methods report Acc-2 and Acc-7 results on MOSEI. However, our dataset only has 5 sentiment categories at most. Thus, we select Acc-2, F1-score and MAE as our metrics, which have been all widely used in SA experiments.

Compared with single-task methods, FGMSA also achieves significant performance improvement in all metrics. This indicates that introducing unimodal sentiment annotations into MSA helps improve performance. Furthermore, some existing methods (e.g., Self-MM) that perform well on public datasets in other domain cannot achieve satisfactory performance on our FMSA-SC. This is because most of the existing MSA datasets do not provide unimodal sentiment annotations. This poses a challenge for exploring modality-specific sentiment cues, which is also a key factor for MSA [60]. Furthermore, these baselines cannot capture fine-grained multimodal sentiments at the phrase level, which hinders their performance.

From the table we can observe that although our proposed model has relatively large FLOPs, its inference speed is very fast. This is because our work uses a multitask framework that incorporates multiple unimodal SA tasks and one MSA task, where the MSA performs phrase-level feature fusion and MIL-based sentiment inference. This makes our model require large FLOPs. However, from the table, we can see that the number of parameters in our models is small. As the phrase-level feature fusion and instance-level inference in our model can be performed in parallel, which makes our models have very fast inference speed. In contrast, we also observe that several baselines using BERT (MULT, BERT-MAG, MISA, SELF-MM, MMIM) have relatively small FLOPs but longer inference times. This is because these baselines take a lot of time to load and run the BERT model, which weakens their applicability to real-time SA tasks.

To verify the effectiveness of our method on MSA datasets of other domains. We also compare our methods with some popular MSA methods on the CH-SIMS, which is a well-known MSA benchmark that contains both multimodal and unimodal sentiment labels for each video. The experimental results for all competing methods on the CH-SIMS are reported in Table V. We can observe that the proposed FGMSA consistently outperforms all baselines in terms of Acc-5, which measures the accuracy of classification results over the 5 sentiment categories, i.e., SN,

TABLE VI
RESULTS (%) OF ABLATION STUDY FOR VERIFYING THE EFFECTIVENESS OF OUR STRATEGIES FOR VISUAL FEATURE FILTERING, MULTIMODAL ATTENTION LEARNING, AND MIL-BASED MSA

Model		Acc-2	Acc-3	Acc-5	F1	MAE
FGMSA-TFN	-full model	**75.71**	57.09	**55.06**	**74.89**	**27.74**
	-w/o vf	68.42	52.23	51.42	66.52	29.46
	-w/o mam	74.90	57.49	54.66	74.42	27.89
	-w/o mil	74.49	**57.89**	53.85	74.46	27.97
FGMSA-LMF	-full model	**75.30**	57.09	53.85	**74.77**	**27.88**
	-w/o vf	68.02	53.04	52.63	63.50	29.56
	-w/o mam	70.45	56.68	54.25	69.85	29.37
	-w/o mil	72.87	**58.70**	**54.66**	73.30	28.13
FGMSA-LF-DNN	-full model	**76.92**	**58.30**	**55.06**	**76.62**	27.41
	-w/o vf	68.83	53.44	52.63	66.23	29.55
	-w/o mam	75.30	57.89	54.66	75.26	**27.22**
	-w/o mil	74.90	57.09	54.25	74.62	27.50

The number in bold indicates the best performance achieved by variants of each model on each metric. For example, the variant {FGMSA-TFN-w/o mil} achieves the best Acc-2 among all variants of FGMSA-TFN. For Acc-2, Acc-3, Acc-5 and F1, the larger, the better. For MAE, the lower, the better.

TABLE VII
(%) RESULTS OF OUR FGMSA IN DIFFERENT CONFIGURATIONS

Tasks	Acc-2	F1	MAE
M	68.83	69.31	30.77
M, T	71.66	71.92	28.17
M, A	70.45	71.15	28.47
M, V	70.41	70.90	28.85
M, T, A	72.06	72.28	28.16
M, T, V	70.45	70.91	28.26
M, A, V	69.23	69.75	28.97
M, T, A, $V^{\dagger}$	70.04	70.14	28.89
M, T, A, V	76.92	76.62	27.41

"M" is the main multimodal task and "T, A, V" are auxiliary unimodal tasks. $V^{\dagger}$ indicates the features of visual modality are not filtered.

WN, NE, WP and SP. This demonstrates the effectiveness and generalization ability of our method.

2) Ablation Study: In this part, we first use the combination of different unimodal tasks to verify the influence of auxiliary unimodal tasks. From Table VII, we can see that FGMSA with partial absence of three unimodal tasks witnesses a performance drop. This indicates that it is insufficient to accurately infer sentiments based on any one or two modalities. This is because speakers may have deceptive facial expressions or use a tone that goes against their true sentiments when discussing stocks. For example, FGMSA with two unimodal tasks witnesses at most 6.51% performance drop on the *Acc-2*.

Then, we test a variant (that is, *M, T, A,* $V^{\dagger}$) of FGMSA that does not use the visual feature filtering method described by (7). From Table VII, we can see that the F1 score of this variant drops dramatically by 6.48% compared to the full model. The main reason for the performance degradation is that video frames may contain noise information irrelevant to sentiments, which, if not filtered, would mislead the SA results. This result demonstrates the effectiveness of our method used for visual feature filtering.

To further validate the effectiveness of our visual feature filtering and phrase-level multimodal feature fusion strategy, we test the performance of FGMSA with the following configurations: i) *FGMSA w/o vf*, which directly uses phrase-level visual features (i.e., $\boldsymbol{f}_v^i$) to learn multimodal features $\boldsymbol{x}_m^i$ via Eq. (8). Here $\boldsymbol{f}_v^i$ can be obtained according to Section III-C. ii) *FGMSA w/o mam*, which can be obtained by removing Multimodal Attention Mechanism (MAM). That is, in Eq. (8) we set $\alpha_{ti} = \alpha_{ai} = \alpha_{vi} = 1$. iii) *FGMSA w/o mil*, which can be obtained by removing the module MIL-based MSA. In this variant, we directly use the average of all phrase-level multimodal features $\boldsymbol{x}_m^i$ to predict the MSA results. The results of the above ablation experiments are shown in Table VI.

As seen from the table, compared to the full model, these variant models witness a performance drop in most of the metrics. For example, the F1 values of all variants have significant performance degeneration. Specifically, variant *FGMSA -w/o vf* suffers the largest performance drop in F1, demonstrating the effectiveness of our strategy for filtering visual features. The effectiveness of the Multimodal Attention Mechanism (MAM) used for feature fusion can be verified by the comparison between FGMSA and variant *FGMSA -w/o mam*. Without the mam, FGMSA cannot adjust its attention on each modality according to interactions between different modality data when predicting sentiments. This makes the variant incorporate the sentiment cues mined from each modality indiscriminately, and thus misleads the sentiment learning. For example, speakers may use irony words (e.g., "great") which if are treated equally with the tones (e.g, sarcastic tone) and facial expressions (e.g., angry expression) would confuse the sentiment predictor. We also observe that variant *FGMSA -w/o mil* has a performance drop. The reason mainly lies in that the average of phrase-level multimodal features leads to the loss of sentiment cues, and thus impedes the performance improvement.

VI. Conclusions and Future Work

In this article, we introduce a new fine-grained MSA dataset built on stock comment videos, namely FMSA-SC. It is the first MSA dataset for the stock market that contains both multimodal and independent unimodal sentiment annotations. Besides, FMSA-SC also provides fine-grained alignment annotations for phrase-level textual, visual and acoustic modality data. This opens the door to future studies of fine-grained MSA for the stock market. In addition, according to the characteristics of FMSA-SC, we design a new fine-grained multimodal multi-task sentiment analysis method, namely FGMSA. It can explore multimodal phrase-level sentiment cues and adaptively suppress noisy information from phrase-level features of each modality. Experimental results show that our FGMSA achieves state-of-the-art performance on FMSA-SC. Furthermore, we also conduct some experiments to verify the effectiveness of our method on MSA datasets of other fields. Specifically, we employ one MSA benchmark dataset (i.e., SIMS) that contains video fragments from movies, TV series, and variety shows, and select several popular MSA methods as competing baselines. The experimental results show that our method consistently outperforms the baseline methods on SIMS.

The high-quality MSA benchmark dataset is the key to improving the performance of MSA tasks on the stock market. However, due to the high annotation cost, our FMSA-SC does not contain large-scale stock comment videos with fine-grained sentiment annotations. To advance the development of large-scale MSA models, we will continue to collect and annotate more high-quality stock comment videos to expand this dataset in the future.

References

[1] M. Baker and J. Wurgler, "Investor sentiment in the stock market," *J. Econ. Perspectives*, vol. 21, no. 2, pp. 129–152, 2007.

[2] A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," Standford Univ., vol. 15, p. 2352, 2012. [Online]. Available: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[3] A. Derakhshan and H. Beigy, "Sentiment analysis on stock social media for stock price movement prediction," *Eng. Appl. Artif. Intell.*, vol. 85, pp. 569–578, 2019.

[4] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Comput. Appl.*, vol. 32, pp. 9713–9729, 2020.

[5] Y. Liu, Z. Qin, P. Li, and T. Wan, "Stock volatility prediction using recurrent neural networks with sentiment analysis," in *Proc. Adv. Artif. Intell. Theory Pract. 30th Int. Conf. Ind. Eng. Appl. Appl. Intell. Syst.*, 2017, pp. 192–201.

[6] R. N. Paramanik and V. Singhal, "Sentiment analysis of indian stock market volatility," *Procedia Comput. Sci.*, vol. 176, pp. 330–338, 2020.

[7] F. Alzazah, X. Cheng, and X. Gao, "Predict market movements based on the sentiment of financial video news sites," in *Proc. IEEE 16th Int. Conf. Semantic Comput.*, 2022, pp. 103–110.

[8] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," *PeerJ Comput. Sci.*, vol. 7, p. e476, 2021, doi: 10.7717/peerj-cs.476.

[9] M. G. Sousa et al., "Bert for stock market sentiment analysis," in *Proc. 31st Int. Conf. Tools Artif. Intell.*, 2019, pp. 1597–1601.

[10] N. Tabari, A. Seyeditabari, T. Peddi, M. Hadzikadic, and W. Zadrozny, "A comparison of neural network methods for accurate sentiment analysis of stock market tweets," in *Proc. ECML-PKDD Workshop Mining Data Financial Appl.*, 2018, pp. 51–65.

[11] T. Zhu et al., "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, vol. 25, pp. 3375–3385, Mar. 16, 2022.

[12] W. Yu et al., "CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[13] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, Nov. 2, 2020.

[14] M. Soleymani et al., "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.

[15] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[16] S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[17] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 26, pp. 529–539, Apr. 17, 2023.

[18] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Sep. 28, 2017.

[19] Z. Fu, F. Liu, Q. Xu, X. Fu, and J. Qi, "LMR-CBT: Learning modality-fused representations with CB-transformer for multimodal emotion recognition from unaligned multimodal sequences," *Front. Comput. Sci.*, vol. 18, no. 4, 2024, Art. no. 184314.

[20] R. Entezari et al., "The role of pre-training data in transfer learning," 2023, *arXiv:2302.13602*.

[21] Y. Zhu, X. Wu, J. Qiang, Y. Yuan, and Y. Li, "Representation learning via an integrated autoencoder for unsupervised domain adaptation," *Front. Comput. Sci.*, vol. 17, no. 5, 2023, Art. no. 175334.

[22] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[23] M. Wen et al., "Large sequence models for sequential decision-making: A survey," *Front. Comput. Sci.*, vol. 17, no. 6, 2023, Art. no. 176349.

[24] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[25] D. Valle-Cruz et al., "Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the H1N1 and the COVID-19 periods," *Cogn. Computat.*, vol. 14, pp. 372–387, 2022, doi: 10.1007/s12559-021-09819-8.

[26] H. K. Sul, A. R. Dennis, and L. Yuan, "Trading on twitter: Using social media sentiment to predict stock returns," *Decis. Sci.*, vol. 48, no. 3, pp. 454–488, 2017.

[27] C.-J. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang, "Financial sentiment analysis for risk prediction," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 802–808.

[28] D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment analysis," *IEEE Trans. Syst. Man Cybernet. Syst.*, vol. 44, no. 8, pp. 1077–1087, Aug. 2014.

[29] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1334–1350, Apr.–Jun. 2021.

[30] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.

[31] S. Poria et al., "Meld: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.

[32] V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 38–45, May/Jun. 2013.

[33] L.-P. Morency, E. Cambria, S. Poria, P. P. Liang, and A. Zadeh, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[34] L. Vadicamo et al., "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 308–317.

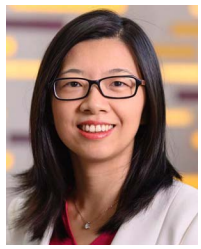
[35] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4351–4360.

[36] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, 2023.
[37] D. Wang et al., "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, Jun. 16, 2022.
[38] W. Guo et al., "LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 1785–1798, Jun. 19, 2020.
[39] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Aug. 29, 2018.
[40] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 11–19.
[41] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[42] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018.
[43] Y. Yang et al., "Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning," *Front. Comput. Sci.*, vol. 18, no. 1, pp. 1–15, 2024.
[44] J. Ma, J. Liu, Q. Chai, P. Wang, and J. Tao, "Diagram perception networks for textbook question answering via joint optimization," *Int. J. Comput. Vis.*, pp. 1–14, Nov. 30, 2023, doi: 10.1007/s11263-023-01954-z.
[45] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Trans. Multimedia Computing, Commun. Appl.*, vol. 19, no. 1, pp. 1–19, 2023.
[46] B. Wang et al., "Non-uniform attention network for multi-modal sentiment analysis," in *Proc. MultiMedia Modeling: 28th Int. Conf.*, 2022, pp. 612–623.
[47] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.
[48] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Inf. Sci.*, vol. 311, pp. 18–38, 2015.
[49] K. Krippendorff, "Association, agreement, and equity," *Qual. Quantity*, vol. 21, no. 2, pp. 109–123, 1987.
[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *North Amer. Chapter Assoc. Comput. Linguistics*, vol. abs/1810.04805, pp. 4171–4186, 2018.
[51] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12449–12460, 2020.
[52] B. Zhang et al., "Wenetspeech: A 10000 hours multi-domain mandarin corpus for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 6182–6186.
[53] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.
[54] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2020, pp. 2359–2369.
[55] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 11, 2014, *arXiv:1412.3555*.
[56] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3866–3870.
[57] F. Sung et al., "Learning to compare: Relation network for few-shot learning," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.
[58] S. Angelidis and M. Lapata, "Multiple instance learning networks for fine-grained sentiment analysis," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 17–31, 2018.
[59] J. Williams, R. Comanescu, O. Radu, and L. Tian, "Dnn multimodal fusion techniques for predicting video sentiment," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018, pp. 64–72.
[60] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[61] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021.
[62] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[63] H. Mao et al., "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2022, pp. 204–213.
[64] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634—5641.

Lingyun Song received the M.S. degree in software engineering and the Ph.D. degree in computer science and technology from Xi'an JiaoTong University, Xi'an, China, in 2014 and 2019, respectively. He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an. His research interests include multimedia analysis, computer vision, data mining, and graph representation learning.

Siyu Chen received the B.S. degree in computer science and technology in 2023 from Northwestern Polytechnical University, Xi'an, China, where he is currently working toward the M.S. degree with the School of Computer Science. His research interests include multimedia analysis, computer vision, and machine learning.

Ziyang Meng received the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2023. His research interests include multimodal sentiment analysis, computer vision, and machine learning.



Mingxuan Sun received the B.S. degree in computer science and engineering from Zhejiang University, Hangzhou, China in 2004, the M.S. degree in computer science from the University of Kentucky, Lexington, KY, USA, in 2006, and the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta, GA, USA, in 2012. She is an Associate Professor with the Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA, USA. Her research interests include trustworthy machine learning, information retrieval, and sequential data mining.

Xuequn Shang is currently a Professor and the Dean of the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. She also is the Deputy Director of CCF Information Systems Committee. Her research interests include multimedia analysis, data mining, machine learning, and biological information.