MDPI

Article

Mixture of Attention Variants for Modal Fusion on Multi-modal Sentiment Analysis

Chao He [1,2], Xinghua Zhang[3], Dongqing Song [1], Yingshan Shen[2], Chengjie Mao[1],Huosheng Wen[4], Dingju Zhu [4] and Lihua Cai [2,4,†]

1 School of Computer Science, South China Normal University, Guangdong Province 510631, China ; chaohe,2019022679,20091223@m.scnu.edu.cn
2 Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangdong Province 528225, China; lee.cai@m.scnu.edu.cn shenyingshan@scnu.edu.cn
3 International United College, South China Normal University, Guangdong Province 528225, China; zhangxinghua@m.scnu.edu.cn
4 School of Software, South China Normal University, Guangdong Province 528225, China; zhudingju,20200666@m.scnu.edu.cn
* Correspondence: L.C., lee.cai@m.scnu.edu.cn; D.Z., zhudingju@m.scnu.edu.cn

Abstract: With popularization of better network access and penetration of personal smartphones in today's world, the explosion of multi-modal data, particularly opinionated video messages, creates urgent demands and immense opportunities for Multi-modal Sentiment Analysis (MSA). Deep learning with attention mechanism has served as the foundational technique for most state-of-the-art MSA models due to its capability in learning complex inter- and intra-relationships among different modalities embedded in video messages, both temporally and spatially. However, Modal fusion is still a major challenge due to vast feature space created by the interactions among different data modalities. To address the modal fusion challenge, we propose a MSA algorithm based on deep learning and attention mechanism, namely Mixture of Attention Variants for Modal Fusion (MAVMF). The MAVMF algorithm includes a two-stage process: in stage one, self-attention is applied to effectively extract image and text features, and the dependency relationships in the contexts of video discourse is captured by a bidirectional gated recurrent neural module; in stage two, four multi-modal attention variants are leveraged to learn the emotional contributions of important features from different modalities. Our proposed approach is end-to-end and has been shown to achieve superior performance to the state-of-the-art algorithms when experimented with two largest public datasets, CMU-MOSI and CMU-MOSEI.

Keywords: multi-modality; attention mechanism; sentiment analysis; feature fusion; deep learning

Citation: He, C.; Zhang, X.; Song, D.; Shen Y.; Mao C.; Wen, H.; Zhu, D.; Cai, L. Mixture of Attention Variants for Modal Fusion on Multi-modal Sentiment Analysis. *Big Data Cogn. Comput.* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. Introduction

The accessibility in 5G networks and popularity of social media have given rise to the ubiquity of opinionated video messages in our cyber world. Amateur users are producing large amount of videos in social media platforms such as TikTok and on the internet to share their sentiments and emotions toward all aspects of daily lives [1,2]. These multi-modal data consist of at least two modalities, commonly including texts, acoustics, and images [3]. For instance, videos on TikTok not only contain creator's spoken language, but also their body movements and facial expressions with pleasing background music and special animation effects.

The rich information in multi-modal data creates immense opportunities for different organizations. The capability in digesting multi-modal data is beneficial to a variety of application scenarios that can greatly enhance user experiences and utility. For example, in autonomous driving, vehicular control unit can leverage cameras to monitor driver's emotions, driving behaviors, and fatigue conditions in real-time, and provide necessary feedback based on the multi-modal signal [4]. This can effectively enhance driving safety

and reduce accidents. In the field of medical and health services, multi-modal information (e.g., patient counseling recordings) can be applied for emotion tendency assessments to assist doctor decisions in patient treatments [5]. Social media platforms can adopt machine learning techniques to automatically poll collective sentiment tendency on videos of a given topic, which can then be relayed to other relevant organizations for decision making [6]. However, with greater possibilities comes greater challenges in leveraging multi-modal data. Unlike single modality data (e.g., texts) for sentiment analysis, the diverse modalities in multi-modal data both complement and interfere with each other, making information fusion extremely challenging.

The challenges in fusing multi-modal data mainly lie in encoding both single modality and cross-modality information, and modeling contextual relationships among the targeted units of analysis (e.g., utterance). Both tasks require significant computing in an enormous feature space, which invalidates manual feature engineering as a solution option. Fortunately, deep learning has shown promising potential in modeling multi-modal data, particularly for Multi-modal Sentiment Analysis (MSA) tasks [7]. Various deep learning methods, including CNN, LSTM/GRU, (Self-)Attention, and BERT, have been leveraged to learn encoding from single and cross modalities [7], and are shown to achieve state-of-the-art performance in MSA tasks [7]. However, there is still no consensus on how to efficiently fuse multi-modal information to remove existing noise while taking contexts into considerations for optimizing performance in MSA.

In this work, we propose a novel algorithm based on a mixture of attention variants for multi-modal fusion in MSA tasks. Our approach divides the multi-modal fusion problem into two stages: 1) we leverage self-attention module to maximize intra-modality information value, and BiGRU module to maintain inter-utterance contexts within each modality; 2) after compressing the feature space using a fully connected module, the resulted tensors from each modality are fed into four different attention variants for multi-modal fusion. In both stages, we apply attention mechanism to distinguish contributions from each modality, assigning higher weights to useful features, while reducing irrelevant background interference.

The main contributions of the current work can be summarized as follows: (1) A comprehensive literature review on both single and multi-modal sentiment analysis is presented. (2) A novel MSA method, namely Mixture of Attention Variants for Modal Fusion (MAVMF) is proposed to solve the multi-modal fusion challenge. (3) Experimental data on two largest benchmark public datasets show that our proposed MAVMF algorithm can effectively extract multi-modal information, and has been shown to demonstrate improvements when compared with other baseline methods.

In the remaining sections, we provide a comprehensive summary on sentiment analysis and new developments in Section 2, introduce our problem definition and proposed algorithm in Section 3, provide the details of the experiments in Section 4, and present the results in Section 5, followed by a discussion in Section 6 and a conclusion in Section 7.

2. Related Work

Before we entered into the video age in social media, text data dominated the Sentiment Analysis (SA) sphere, and were considered the default data modality in SA. However, research in SA using only text data could suffer from issues like "emotional gap" and "subjective perception" [3], leading to unsatisfactory results in SA tasks. Compared with text-only data, user-recorded videos convey emotional information through subscripts (e.g., texts), images, and acoustic signals embedded in them. The popularity of video data gave rise to MSA, in which various modalities are leveraged to corroborate each other for better recognition performance [8]. Since effective extraction of features within single modalities serves as the foundation and prerequisite for MSA tasks, it is necessary to use feature extractors to extract internal features within each modality. In addition, since the fusion of multi-modal features is critical to the success of MSA tasks, effective fusion algorithms is

at the core of MSA research. From these two perspectives, we will cover both single- and multi-modality SA in this section.

2.1. Single Modality Sentiment Analysis

2.1.1. Text Sentiment Analysis

In Text Sentiment Analysis (TSA), we aim to discover the emotional attitudes expressed by the authors through analyzing the emotions within the text. Before the advent of text analysis technology, people had to manually read and analyze the emotions conveyed in the text, resulting in a significant increase in workload. In addition, manual classification is prone to human errors. Therefore, using automation technology to infer text sentiments will significantly improve label efficiency in TSA.

TSA tasks can be categorized into word-level, sentence-level, and document-level inferences. [9] Document-level SA focuses on the overall emotional tendency, which is obtained by assigning different sentiment contributions to different sentences, and aggregating the sentiment tendencies of all sentences in the document. Sentence-level SA focuses on each individual sentence in a document and studies the emotional polarity of the sentence based on the sentiment contributions of the words within the sentence. Word-level SA focuses on each word that appears in a sentence and directly determines its emotional polarity through sentiment lexicons.

All TSA methods can be simply divided into rule-based and machine-learning-based methods [10]. The rule-based approaches use pre-designed rules, such as sentiment lexicons, to determine text sentiments. For example, sentiment lexicons may define the polarity scores of emotion words, and the overall sentiment polarity is determined by aggregating the positive and negative scores of the words. The one with the higher score is selected as the final sentiment polarity. The performance of rule-based SA methods largely rely on the accuracy of the scoring on each word and the comprehensiveness of the lexicon set. Due to its simplicity in implementation, rule-based SA methods are widely adopted by researchers and practitioners [11]. For instance, Thelwall et al. [12] proposed the SentiStrength algorithm; Saif et al. [13] developed the SentiCircles platform for SA on Twitter; Li et al. [14] constructed a lexicon to effectively enhance sentiment analysis performance; Kanayama et al. [15] proposed a syntactic-based method for detecting sentiment polarity; while Rao et al. [16] conducted sentiment analysis based on document topic classification.

The machine-learning-based SA methods aim to automate SA tasks using supervised models. For example, Chen et al. [17] proposed a novel SA algorithm to extract sentiment features from mobile app reviews and used Support Vector Machines (SVM) for sentiment classification. Zhao et al. [18] used supervised algorithms to perform binary classification on product review data, classifying the comments into positive and negative categories. Kiritchenko et al. [19] proposed a SVM algorithm for short and informal texts. Silva et al. [20] applied ensemble learning by combining various classifiers such as random forests and SVMs. More recently, deep learning approaches have emerged as a new avenue of research in TSA. Kim et al. [21] used Convolutional Neural Networks (CNN) for SA, and was able to demonstrate excellent performance. Makoto et al. [22] combined spatial pyramid pooling with max pooling, and used gated neural networks to classify user review texts. Meng et al. [23] proposed a multi-layer CNN algorithm, and was able to prove its superiority through experiments. Jiang et al. [24] combined Long Short Term Memory (LSTM) networks [25] with CNNs to handle the dependency on distant sentences. Luo et al. [26] introduced a gated Recurrent Neural Network (RNN) to enhance the contextual relationships between words and texts. Minh et al. [27] proposed three variants of neural networks [28] to capture the long-term dependencies of information.

2.1.2. Image Sentiment Analysis

Image Sentiment Analysis (ISA) mainly focuses on modeling users' facial expressions and postures contained in an image to infer their emotional tendencies. Colombo et al. [29] first proposed an automatic emotion retrieval system that effectively extracts image features

and performs emotion classification. Singh et al. [30] applied CNN with domain specific fine tuning to classify sentiments on Flickr images. Yang et al. [31] created a learning framework that explores only affective regions in an image and combined it with a CNN to classify sentiment for an image. Yang et al. [32] proposed a weakly supervised coupled CNN with two branches to leverage localized information of an image for ISA. Kumar et al. [33] constructed a visual emotion framework for emotion feature extraction using the Flickr dataset. Truong et al. [34] developed item-oriented and user-oriented CNN to better capture the interaction of image features with specific expressions of users or items for inference of user review sentiments. You et al. [35] extracted features from local image regions and conducted ISA by incorporating an attention mechanism into the proposed network. Wu et al. [36] proposed a scheme for ISA that leverage both the inference on the whole image and sub-images that contained salient objects. Zheng et al. [37] introduced an "Emotion Region Attention" module, while Li et al. [38] proposed a novel SentiNet model for ISA.

2.1.3. Speech Emotion Analysis

Compared to text and image SA tasks, the development of Speech Emotion Analysis (SEA) has been relatively slow. SEA focuses on analyzing the emotions based on factors such as tone, bandwidth, pitch, and duration of user speech [39]. Since deep learning techniques have been shown to improve speech recognition performance [40], researchers have proposed various neural network-based models to enhance the accuracy of speech emotion recognition [41,42].

2.2. Multi-modal Sentiment Analysis

Existing and emergent social media platforms have enabled common users to post self-recorded videos to share their day-to-day living experiences and sentiments on any subjects. This leads to an explosion of multi-modal information on the internet, and creates tremendous opportunities for MSA [43]. Morency et al. [1] created the YouTube dataset and constructed a joint model to extract multi-modal features for SA. Poria et al. [44] applied single modality feature extractors (e.g., CNN on text embeddings and Part-of-Speech taggings) on the visual, audio, and textual channels, and trained a multi-kernel learning classifier for MSA. Zadeh et al. [2] introduced a multi-modal lexicon to better capture the interactions between facial gestures and speech. They also published the CMU-MOSI [45] dataset, which became the first benchmark dataset to support research in MSA. Zadeh et al. [46] presented a tensor fusion network model, which learns the interactions within and between text, vision, and acoustics channels. Chen et al. [47] proposed a novel SA model, which comprises a gated multi-modal embedding module for information fusion in noisy environments, and an LSTM module with temporal attention for higher-resolution word-level fusion.

The aforementioned works only consider the fusion of information between modalities without considering the dependencies between contexts. In order to improve the performance of MSA, Poria et al. [48] introduced an LSTM-based framework that captures the mutual dependencies between utterances using contextual information. In another work, Poria et al. [49] proposed a user-opinion-based model that combines the three modality inputs using a multi-modal learning approach. Zadeh et al. [50] proposed multiple attention blocks to capture information from the three modalities.

More recently, Wang et al. [51] proposed a novel method Text Enhanced Transformer Fusion Network (TETFN) that learns text-oriented pairwise cross-modal mappings for obtaining effective unified multimodal representations. Yang et al. [52] applied BERT to translate visual and audio features into text features to enhance the quality of both visual and audio features. Wu et al. [53] extracted bimodal features from the acoustic-visual, acoustic-textual, and visual-textual pairs with multi-head attention module to improve video sentiment analysis tasks. Wang et al. [54] created a lightweight Hierarchical Crossmodal Transformer with Modality Gating (HCT-MG) for MSA through primary

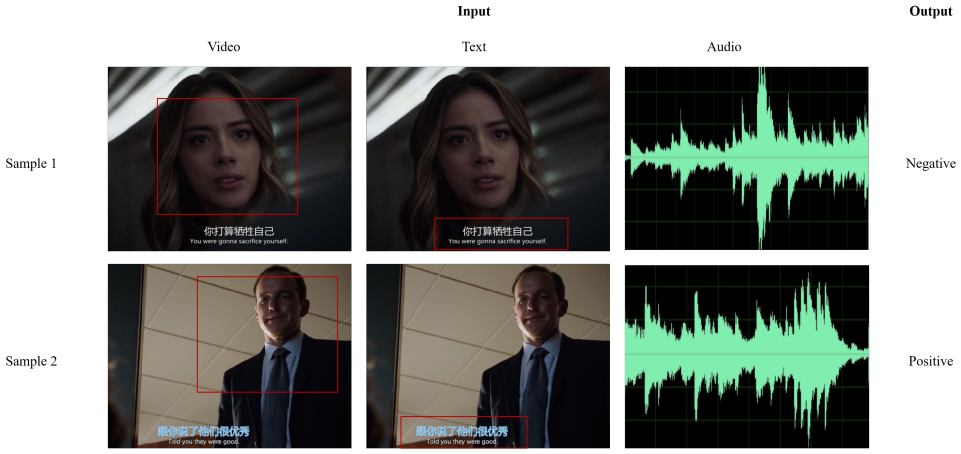

Figure 1. Selected samples from CMU-MOSI dataset. Each sample represents an utterance from the given video. The output labels are obtained by using our proposed MAVMF algorithm.

modality identification. He et al. [55] proposed the Multimodal Temporal Attention (MMTA) algorithm, which considers the temporal effects of all modalities on each unimodal branch to balance the interaction between unimodal branches and adaptive intermodal balance. Mai et al. [56] leveraged the contrastive learning framework both within modality and between modalities for the MSA tasks.

Although these scholars have achieved promising results, there is still room for improvement. In SA tasks, Within modality representations are as important as inter-modality information fusion. The aforementioned research methods do not fully address both the extraction of modality-specific features and the fusion of information between modalities. Particularly, the interaction and fusion of multi-modal features may lead to redundant information in the target network, making it challenging to focus on important information. Therefore, it is critical to identify the contributions of features from different modalities to SA at each stage of the deep learning network, which is the goal of our proposed MAVMF algorithm.

3. Method

3.1. Problem Definition

In this work, videos are considered the source of multi-modal data for sentiment analysis. A video usually contains a series of consecutive image frames, and user's emotional tendency could be different or related in consecutive frames. Because of this, the video is processed into video utterances, each of which contains the same emotional tendency of the user as shown in Fig 1. We aim to perform sentiment analysis on the utterance level in a given video.

Assuming a dataset contains m videos, $D = [V_1, V_2, \ldots, V_m]$. For the i-th video, V_i, the video is composed of n_i video segments or utterances, $V_i = [u_{i1}, u_{i2}, \ldots, u_{in_i}]$, where u_{i1} denotes the first utterance in V_i, and n_i denotes the total number of utterances in V_i. Each utterance $u_{ij}, 1 \leq i \leq m$ and $1 \leq j \leq n_i$, contains a feature vector $u_{ij} = [t_{ij}, v_{ij}, a_{ij}]$ representing three modalities, i.e., t_{ij} is the text feature representation, v_{ij} is the visual feature representation, and a_{ij} is the audio feature representation.

Assuming there are C classes of emotional categories for the user's video, the goal is to label the emotional category of each video utterance. In order to perform sentiment classification, the utterances in the video, except for u_{ij}, is considered as the context of u_{ij}, and the accuracy and F-1 score are used as the evaluation metrics for the model.

3.2. Model

The MAVMF algorithm can be divided into five steps from an end-to-end pipeline as shown in Fig 2: (A) Single-modal feature representation, (B) Single-modal attention, (C)

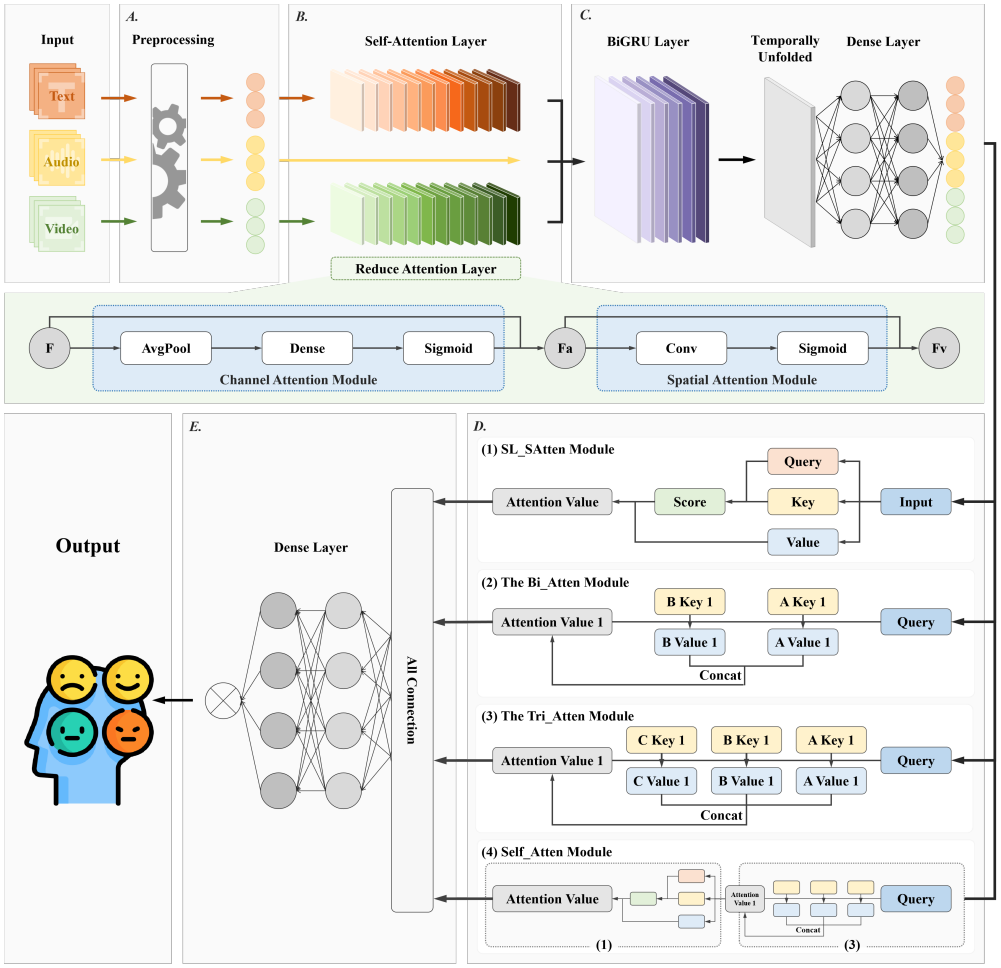

Figure 2. MAVMF Architecture. A - preprocessing input multi-modal data. B - single modality self-attention and reduce-attention module for text and visual modalities. C - BiGRU + Fully connected dense layer(s). D - four attention and self-attention modules for modal fusion. E - concatenation + fully connected dense layer(s) with softmax activation function for prediction.

Contextual feature extraction, (D) Multimodal feature fusion, and (E) Sentiment classification. The concrete architecture of MAVMF is illustrated in Algorithm 1. We employ $\odot$ for dot product, $\otimes$ for element-wise multiplication, and $\oplus$ for feature concatenation in all formulas in the below sections.

3.2.1. Single-modal feature representation

Due to the distinct semantic spaces of text, image, and audio, different feature extractors should be used for extracting features within each modality. We adopt the following single-modal feature extractors that were chosen by other studies with the CMU-MOSI and CMU-MOSEI datasets. Specifically, for the CMU-MOSI dataset, we use the utterance-level features provided by Poria et al. [48] as inputs to the MAVMF model; for the CMU-MOSEI dataset, we employ the CMU-Multi-modal Data SDK [50] tool to extract corresponding text, audio, and video features as inputs to the MAVMF model.

3.2.2. Single-modal attention

The attention mechanism enables the target network to prioritize high-contributing features and disregard interference from background information. This section primarily focuses on the single-modal attention modules, which employ distinct attention mechanisms to encode image and text features independently, enhancing the extraction of single-modal features. The Single-modal attention module comprises two components: the Reduced

Algorithm 1: MAVMF Architecture

Input : text t, audio a, visual v of train data U and test data R
Output: predictions for R
Procedure TRAIN(U,R):
  Unimodal:
    Train $BiGRU$ and Dense layers with t, a, and v
    $t \longleftarrow Self\_Atten(t)$
    $X_t \longleftarrow getBiGRUFeatures(t)$
    $D_t \longleftarrow Dense(X_t)$
    $v \longleftarrow RAtten(v)$
    $X_v \longleftarrow getBiGRUFeatures(v)$
    $D_v \longleftarrow Dense(X_v)$
    $X_a \longleftarrow getBiGRUFeatures(a)$
    $D_a \longleftarrow Dense(X_a)$
  Multimodal:
    Fusion text, audio and visual features
    $SL\_SAtten_t \longleftarrow SL\_SAtten(D_t)$
    $SL\_SAtten_v \longleftarrow SL\_SAtten(D_v)$
    $SL\_SAtten_a \longleftarrow SL\_SAtten(D_a)$
    $Bi\_Atten_{ta} \longleftarrow Bi\_Atten(concat(D_t, D_a))$
    $Bi\_Atten_{tv} \longleftarrow Bi\_Atten(concat(D_t, D_v))$
    $Bi\_Atten_{av} \longleftarrow Bi\_Atten(concat(D_a, D_v))$
    $Tri\_Atten_{tva} \longleftarrow Tri\_Atten(concat(D_t, D_v, D_a))$
    $Self\_Atten_{tva} \longleftarrow selfattention(Tri\_Atten_{tva})$
  Fuse feature and classification:
    $D \longleftarrow concat(D_t, D_v, D_a)$
    $O_{SL\_SAtten} \longleftarrow concat(SL\_SAtten_t, SL\_SAtten_v, SL\_SAtten_a)$
    $O_{Bi\_Atten} \longleftarrow concat(Bi\_Atten_{ta}, Bi\_Atten_{tv}, Bi\_Atten_{av})$
    $O_{out} \longleftarrow concat(D, O_{SL\_SAtten}, O_{Bi\_Atten}, Tri\_Atten_{tva}, Self\_Atten_{tva})$
    $output \longleftarrow softmax(O_{out})$
Procedure TEST(R):
  R passed through the learned model to get the results
  $Y \longleftarrow MAVMF(t, a, v)$

Attention ($RAtten$) block for the visual modality and the Self-Attention ($Self_Atten$) block for the text modality. The structure of the Single-modal attention module is illustrated in Figure 2 (B).

The RAtten Block. The RAtten block employs Channel Attention and Spatial Attention to encode internal features of the image, enhancing features that significantly contribute to emotions while suppressing background features. The image's feature vector sequentially passes through the Channel Attention and Spatial Attention modules, capturing the importance of each channel and feature map in the image. The specific process can be expressed as:

$$F_a = F \otimes \sigma(Dense(AvgPool(F))) \tag{1}$$

$$F_v = F_a \otimes \sigma(Conv(F_a)) \tag{2}$$

F is the image information extracted by the preprocessing method described in Section 3.2.1, $AvgPool$ is one-dimensional global average pooling, $Dense$ represents fully connected layer, σ represents $Sigmoid$ function, F_a represents the output of the image features F after passing through the channel attention, $Conv$ represents a one-dimensional convolution operation, and F_v represents the output of the spatial attention.

The Self_Atten Block. The Self_Atten block applies self-attention mechanism to encode text features, which can take into account the mutual influence among video segments. Since words in the same sentence in a video have different semantic associations, the self-attention mechanism can calculate the semantic association between a word in a sentence and other words in the same sentence, providing them different weights. This process can be expressed as follows:

$$m_i = x_i \odot x_i^T \tag{3}$$

$$n_i = softmax(m_i) \tag{4}$$

$$o_i = n_i \odot x_i \tag{5}$$

$$a_i = o_i \otimes x_i \tag{6}$$

$$t_i = x_i \oplus a_i \tag{7}$$

x_i represents the text feature vectors extracted by the preprocessing method described in Section 3.2.1, a_i is the output after self-attention and represents the importance of different words in each utterance, t_i is the output of the text features after passing through the Self_Atten block.

3.2.3. Contextual Feature Extraction

In order to capture the contexts and dependencies between utterances in each modality, we feed the single-modal attention features extracted in Section 3.2.2 for both the visual and text modalities, and the preprocessed acoustic features to a *BiGRU* module separately. The *BiGRU* module consists of two Gated Recurrent Units (GRUs) with opposite directions, which can effectively capture the spatio-temporal information between video clip sequences, and can also capture the forward and backward long-term dependencies between video clip sequences. The working principles of *BiGRU* can be expressed as follows:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{8}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{9}$$

$$\tilde{h_t} = tanh(r_t * Uh_{t-1} + Wx_t + b_h) \tag{10}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h_t \tag{11}$$

x_t is the input feature sequence of the current node, h_{t-1} is the hidden layer state of the previous GRU unit of the current node, r_t and z_t are the reset gate and update gate of the GRU unit, W_r , b_r , W_z , b_z , U_r and U_z are the weight parameters of the target network, σ is the corresponding *sigmoid* function of the network, $*$ represents the multiplication of the corresponding feature vectors.

To better handle the heterogeneity within each modality, and fully explore the internal correlations of single-modal features, the representation of each modality obtained from the above BiGRU module is unfolded temporally and fused through fully connected dense layer(s). This can be expressed as follows:

$$D_t = tanh(W_t B_t + b_t) \tag{12}$$

$$D_a = tanh(W_a B_a + b_a) \tag{13}$$

$$D_v = tanh(W_v B_v + b_v) \tag{14}$$

B_t, B_a, and B_v are the output features of text, audio, and video after going through the *BiGRU* module, W_t, b_t, W_a, b_a, W_v, b_v are the weight parameters of the target network, $D_t \in R^{u \times d}$, $D_a \in R^{u \times d}$, $D_v \in R^{u \times d}$ are the fully connected layer's text, acoustic, and visual information, u represents the total number of sentences, d represents the number of neurons in the fully connected layer.

3.2.4. Multimodal Feature Fusion

This module comprises attention modules at four different levels and dimensions: (1) sentence-level self-attention module; (2) bi-modality attention module; (3) tri-modality attention module; and (4) self-attention module on (3). It is based on the outputs from Section 3.2.3.

(1) **SL_SAtten Module**. The contribution of internal features in each modality to users' emotional tendency is often different. For example, in the sentence "The weather is really good today! I really like this kind of weather", the word "like" contributes more to the user's emotions than the word "weather". Therefore, we propose a sentence-level self-attention mechanism, referred to as $SL-SAtten$, to select the emotional contribution of words within the modality from the sentence level. Taking the text modality as an example, assume there are u sentences in total in the text modality. For each sentence x_i where $1 \leq i \leq u$, the working principle of the SL_SAtten module is as follows:

$$m_i = x_i \odot x_i^T \tag{15}$$

$$n_i = softmax(m_i) \tag{16}$$

$$o_i = n_i \odot x_i \tag{17}$$

$$a_i = o_i \otimes x_i \tag{18}$$

$$T = a_1 \oplus a_2 \ldots \oplus a_u \tag{19}$$

$$O_{SL-SAtten} = T \oplus V \oplus A \tag{20}$$

Here, a_i is the output of discourse x_i after self-attention, indicating the importance of different words in each utterance. Then we concatenate all the outputs to get the corresponding text feature T. In the same vein, you can get the corresponding visual feature V and the corresponding audio feature A. By concatenating the text, visual, and audio features, you can get the final output of the SL_SAtten module O_{SL_SAtten}.

(2) **The Bi_Atten Module**. In order to improve the interaction between pairs of modalities in video data, a bi-modality attention is proposed. This module aims to integrate two different modalities from different semantic spaces by enhancing the connections between them while eliminating the interference from background information in them, and thus the learning will be able to focus on the associations between them. Take text and visual modalities as an example, suppose D_t and D_v are the output feature vectors of the text and visual modalities after going through the context feature extraction module, its working principle is as follows:

$$m_1 = D_t \odot D_v^T \tag{21}$$

$$n_1 = softmax(m1) \tag{22}$$

$$o_1 = n_1 \odot D_v \tag{23}$$

$$a_1 = o_1 \otimes D_t \tag{24}$$

$$m_2 = D_v \odot D_t^T \tag{25}$$

$$n_2 = softmax(m_2) \tag{26}$$

$$o_2 = m_2 \odot D_t \tag{27}$$

$$a_2 = o_2 \otimes D_v \tag{28}$$

$$O_{BC-Atten}(vt) = a_1 \oplus a_2 \tag{29}$$

$O_{BC\_Atten}(vt)$ is the output feature vector fused from video and text modalities, which can be used for subsequent emotion classification. In similar ways, we can obtain $O_{BC\_Atten}(av)$ for acoustic and visual, and $O_{BC\_Atten}(at)$ for acoustic and text modalities.

(3) **The Tri_Atten Module**. In order to model the interactions among all three modalities in video, a tri-modal attention is proposed. Assuming the text, visual, and acoustic feature vectors after the context feature extraction module are D_t, D_v, and D_a respectively, we first concatenate and fuse the text and image, text and acoustic, image and acoustic modality information, then use a fully connected network to map the information to the same semantic space, thereby initially fusing the information between different modality pairs. The process is shown in the following formulas:

$$F_{TV} = tanh((D_T \oplus D_V)W_{tv} + b_{tv}) \tag{30}$$

$$F_{TA} = tanh((D_T \oplus D_A)W_{ta} + b_{ta}) \tag{31}$$

$$F_{AV} = tanh((D_A \oplus D_V)W_{av} + b_{av}) \tag{32}$$

W_{tv}, W_{ta}, W_{av}, b_{tv}, b_{ta}, b_{av} are the weights and biases of the fully connected layer, and $F_{TV}, F_{TV}, F_{TV} \in R^{u \times d}$ is the pairwise fused feature vectors, where d is the number of neurons in the fully connected layer.

In order to further extract effective features, the feature vector of the third modality is multiplied by the results of the pairwise fused feature vectors obtained in euqations 30,31,32 to produce matrice $C_k(k = 1, 2, 3)$. Then the $softmax$ function is used to calculate the attention distribution of the feature vector fusion results $P_k(k = 1, 2, 3)$, forming tri-modal attention $T_k(k = 1, 2, 3)$, and finally we obtain the tri-modal fusion information Tri_{ATV},Tri_{VTA},Tri_{TAV} through matrix multiplication operations. This is then concatenated to form a feature vector O_{Tri_Atten}, which is the output of the Tri_Atten module. The process is shown in the following formulas:

$$C_1 = F_A \odot F_{TV}^T \tag{33}$$

$$C_2 = F_V \odot F_{TA}^T \tag{34}$$

$$C_3 = F_T \odot F_{AV}^T \tag{35}$$

$$P_1 = softmax(C_1) \tag{36}$$

$$P_2 = softmax(C_2) \tag{37}$$

$$P_3 = softmax(C_3) \tag{38}$$

$$T_1 = P_1 \odot F_A \tag{39}$$

$$T_2 = P_2 \odot F_V \tag{40}$$

$$T_3 = P_3 \odot F_T \tag{41}$$

$$Tri_{ATV} = T_1 \otimes F_{TV} \tag{42}$$

$$Tri_{VTA} = T_2 \otimes F_{TA} \tag{43}$$

$$Tri_{TAV} = T_3 \otimes F_{AV} \tag{44}$$

$$O_{Tri\_Atten} = Tri_{ATV} \oplus Tri_{VTA} \oplus Tri_{TAV} \tag{45}$$

$C_1, C_2, C_3 \in R^{u \times u}$, $T_1, T_2, T_3 \in R^{u \times d}$, and $O_{Tri\_Atten} \in R^{u \times 3d}$.

(4) **Self_Atten Module**. The output of the Tri_Atten module may carry redundant features. To filter out redundant information, we apply a self-attention module for feature selection. This process is shown in the following formulas:

$$m_i = O_{Tri\_Atten} \odot O_{Tri\_Atten}^T \tag{46}$$

$$n_i = softmax(m_i) \tag{47}$$

$$o_i = n_i \odot x_i \tag{48}$$

$$a_i = o_i \otimes x_i \tag{49}$$

$$O_{Self\_Atten} = x_i \oplus a_i \tag{50}$$

O_{Self_Atten} is the output feature vector of the Self_Atten module.

3.2.5. Multimodal Sentiment Classification

Finally, the Multimodal Sentiment Classification module concatenates and combines the feature vectors obtained above, and uses a fully connected layer to integrate and classify sentiments based on both inter-modal and intra-modal information. It is shown as follows:

$$\begin{aligned} out = {} & D_v \oplus D_t \oplus D_a \\ & \oplus O_{SL-SAtten} \oplus O_{CS-SAtten} \\ & \oplus O_{BC-Atten}(vt) \oplus O_{BC-Atten}(at) \\ & \oplus O_{BC-Atten}(av) \oplus O_{Self-Atten} \end{aligned} \tag{51}$$

$$output = softmax(out) \tag{52}$$

$output$ is the final output information of the MAVMF model.

4. Experiments

The experiments are conducted on a Windows system, using an NVIDIA GeForce RTX 2060 graphics card with 8G running memory. Python is used as the programming language with the Keras framework. The effectiveness of the MAVMF model was validated on the two benchmark datasets, CMU-MOSI and CMU-MOSEI.

4.1. Data

(1) The CMU-MOSI dataset includes 93 videos sourced from YouTube, covering topics such as movies, products, and books. There are a total of 2199 utterances, each of which has been labeled as positive or negative. The experiment uses training and test sets of 62 and 31 videos respectively.

(2) The CMU-MOSEI dataset includes 3229 videos, with a total of 22676 utterances, each with an emotional score in the range of $[-3, +3]$. For the purpose of sentiment classification, utterances with a score greater than or equal to 0 are labeled as positive, while those with scores less than 0 are labeled as negative. The experiment uses training, test, and validation sets of 2250, 679, and 300 videos respectively. Detailed information about the CMU-MOSI and CMU-MOSEI datasets are shown in Table 1:

Table 1. The details of the CMU-MOSE and CMU-MOSEI datasets.

Description	CMU-MOSI		CMU-MOSEI	
	Training Set	Test Set	Training Set	Test Set
# Video	62	31	2250	679
# Utterance	1447	752	16216	4625
# Pos Utterance	709	467	11498	3281
# Neg Utterance	738	285	4718	1344

From the detailed information about the CMU-MOSE and CMU-MOSEI datasets in Table 1, it is apparent that the number of positive utterance samples in these two datasets is greater than the number of negative utterance samples, leading to an imbalance in the distribution of positive and negative samples. Therefore, accuracy and F1 scores are used as evaluation metrics for the models.

4.2. Parameter Tuning

During the experiments, we investigate the impacts of different learning rates and batch sizes on model performance. The learning rates chosen are 0.05, 0.01, 0.005, and 0.001, and the batch sizes chosen are 32 and 64. The parameters that yielded the best results are

used for the final model. The final parameter settings for the MAVMF model are shown in Table 2:

Table 2. Experimental parameter settings.

Parameter	Value
BiGRU unit	300
BiGRU Dropout	0.5
fully connected unit	100
fully connected Dropout	0.5
activation function	tanh
learning rate	0.001
batch processing	32
number of iterations	64
optimization function	Adam
loss function	categorical crossentropy

4.3. Baseline Models

To compare the performance of the MAVMF model in the MSA tasks, for the CMU-MOSI dataset, we use the following baseline methods:
(1) GME-LSTM [47]: This model is composed of two modules. One is the gated multimodal embedding module, which can perform information fusion in noisy environments. The other is an LSTM module with temporal attention, which can perform word-level fusion with a higher fusion resolution.
(2) MARN [50]: This model captures the interrelationship of text, images, and speech in time series through multiple attention modules and stores the information in a long short-term hybrid memory.
(3) TFN [46]: This model encodes intra-modal and inter-modal information through embedding sub-networks within a single modality and tensor fusion strategy.
(4) MFRN [57]: This model first stores the modality information through a long short-term fusion memory network, fully considering the information of other modalities when encoding a certain modality, thereby enhancing modality interactivity. Then it further considers the information of other modalities when encoding a single modality through a modality fusion recurrent network. Finally, further information fusion is achieved through the attention mechanism.
(5) Multilogue-Net [58]: Based on a recurrent neural network, this model captures the context of utterances and the relevance of the current speaker and listener in the utterance through multi-modal information.
(6) DialogueRNN [59]: This model tracks the state of independent parties throughout the dialogue process and processes the information through a global GRU, party GRU, and emotion GRU units, and using it for emotion classification.
(7) AMF-BiGRU [60]: This model first extracts the connections between contexts in each modality through BiGRU, then merges information through cross-modal attention, and finally uses multimodal attention to select contributions from the merged information.

For the CMU-MOSEI dataset, we have the following baseline methods:
(1) MFRN [57]: same as above.
(2) Graph-MFN [61]: This model's concept is similar to the MFN model, except that Graph-MFN uses a dynamic fusion graph to replace the fusion block in the MFN model.
(3) CIM-Att [62]: This model first uses BiGRU to extract the intra-modal context features, then inputs these context features into the CIM attention module to capture the associations between pairwise modalities, and then concatenates the context features and CIM module features for sentiment classification.
(4) AMF-BiGRU [60]: same as above.

Table 3. Comparison of performance on different models.

Network Model	CMU-MOSI	
	Accuracy (%)	F-1
GME-LSTM [47]	76.50	73.40
MARN [50]	77.10	77.00
TFN [46]	77.10	77.90
MFRN [57]	78.10	77.90
Multilogue-Net [58]	81.19	80.10
DialogueRNN [59]	79.80	79.48
AMF-BiGRU [60]	82.05	82.02
MAVMF	**82.31**	**82.20**

Table 4. Comparison of performance on different models.

Network Model	CMU-MOSEI	
	Accuracy (%)	F-1
MFRN [57]	77.90	77.40
Graph-MFN [61]	76.90	77.00
CIM-Att [62]	79.80	77.60
AMF-BiGRU [60] [64]	78.48	78.18
MAM [63]	81.00	78.90
MAVMF	**81.10**	**79.48**

(5) MAM [63]: This model first uses CNN and BiGRU to extract features from text, speech, and image signals, then applies cross-modal attention and self-attention for information fusion and contribution selection.

5. Results

5.1. CMU-MOSI

Table 3 presents a comparison of the experiments between the MAVMF model and the chosen baseline models on the CMU-MOSI dataset. The MAVMF model shows some improvement in both classification accuracy and F-1 score. Specifically, the accuracy of the MAVMF model has increased by 5.81%, 5.21%, 5.21%, 4.21%, 1.12%, 2.51%, 0.26% when compared to the GME-LSTM, MARN, TFN, MFRN, Multilogue-Net, DialogueRNN, and AMF-BiGRU models, respectively. The F1 score of the MAVMF model has increased by 8.8%, 5.2%, 4.3%, 4.3%, 2.1%, 2.36%, 0.18% when compared to the GME-LSTM, MARN, TFN, MFRN, Multilogue-Net, DialogueRNN, and AMF-BiGRU models, respectively.

5.2. CMU-MOSEI

Table 4 presents a comparison of the experiments between the MAVMF model and the chosen baseline models on the CMU-MOSEI dataset. The MAVMF model shows some improvement in both classification accuracy and F-1 score. Specifically, the accuracy of the MAVMF model has increased by 3.2%, 4.2%, 1.3%, 2.26%, 0.1% when compared to the MFRN, Graph-MFN, CIM-Att, AMF-BiGRU, and MAM models, respectively. The F-1 score of the MAVMF model has increased by 2.08%, 2.48%, 1.88%, 1.3%, 0.58% when compared to the MFRN, Graph-MFN, CIM-Att, AMF-BiGRU, and MAM models, respectively.

5.3. Modality Analysis

In order to further analyze the impacts of features from different modalities on classification performance of the MAVMF model, experiments were conducted on the CMU-MOSI dataset for both bi-modal and tri-modal feature sets. The experimental results are shown in Fig 3. We use T, V, and A to represent the text, visual, and acoustic modalities, respectively.

From the above results, when compared to the selected baseline models, the classification accuracy of text plus acoustic modalities has increased by $0.53\% - 1.09\%$, the

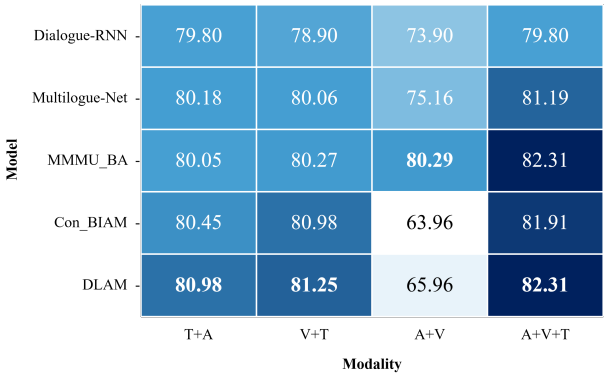

Figure 3. Visualization of the impacts of modalities on the performances of the baseline algorithms and MAVMF.

classification accuracy of text plus visual modalities has increased by 0.27%-2.35%, and the classification accuracy of acoustic plus visual plus text has increased by $0.4\% - 2.51\%$. Apart from the fusion results of video and acoustic modalities, the MAVMF model has achieved the best performance in all other modality fusion methods. The fusion result of acoustic and visual features is the worst, reflecting that the emotional expression polarity of acoustic and visual modalities is weaker than text, and that they may be affected by background noise. This is consistent with the experimental results in the literature [65]. In addition, in MSA tasks, the classification performance of fusing all three modalities is the best, which proves the necessity of leveraging multi-modal information in SA.

5.4. Ablation Study

To understand the impacts of the different modules applied in the MAVMF model, we conduct experiments on variants of the MAVMF model using the CMU-MOSI dataset, and analyze the experimental results. We include the following MAVMF variant models:

(1) MAVMF_Concat: it includes Module A and E as shown in Figure 2. (2) MAVMF_SAtten: It includes Module A, the self-attention module in Module B for text modality, and Module E, as shown in Figure 2. (3) MAVMF_RAtten: It includes Module A, the reduce-attention module in Module B for visual modality, and Module E, as shown in Figure 2. (4) MAVMF_RAtten_SAtten: It includes Module A, Module B, and Module E, as shown in Figure 2. (5) MAVMF_BiGRU: It includes Module A, Module B, Module C, and Module E, as shown in Figure 2. (6) MAVMF_SL-SAtten: It includes Module A, Module B, Module C, plus a sentence level self-attention module, and Module E, as shown in Figure 2. (7) MAVMF_Bi-Atten: It includes Module A, Module B, Module C, a sentence level self-attention module, a dual modality cross modal attention module, and Module E, as shown in Figure 2. (8) MAVMF_Tri-Atten : It includes Module A, Module B, Module C, a sentence level self-attention module, a dual modality cross modal attention module, a three modality cross modal attention module, and Module E, as shown in Figure 2. (9) MAVMF_Self-Atten: It includes Module A, Module B, Module C, Module D, and Module E, as shown in Figure 2.

Table 5 compares the proposed MAVMF model with its variant on the CMU-MOSI dataset. From the experiments, we see that the multi-modal sentiment classification accuracy of the MAVMF model gradually improves after adding each module. The text self-attention module, visual reduced attention module, single modality attention module, bidirectional gated recurrent unit module, sentence-level self-attention module, dual modality cross-modal attention module, three-modality cross-modal attention module, and self-attention module respectively contribute 2.27%, 2.53%, 4.26%, 5.32%, 0.4%, 0.4%, 0.26%, 0.4%, and 0.53% in classification accuracy. The marginal improvements get smaller as the complexity of the model increase.

Table 5. Comparison of performance on different models.

Network Model	CMU-MOSI	
	Accuracy (%)	F-1
MAVMF_Concat	70.74	71.01
MAVMF_SAtten	73.01	73.05
MAVMF_RAtten	73.27	73.24
MAVMF_RAtten_SAtten	75.00	74.95
MAVMF_BiGRU	80.32	80.21
MAVMF_SL-SAtten	80.72	81.09
MAVMF_Bi-Atten	81.12	81.10
MAVMF_Tri-Atten	81.38	81.41
MAVMF_Self-Atten	81.78	82.06
MAVMF	**82.31**	**82.20**

6. Discussion

Multi-modal sentiment analysis tasks are commonplace in a diverse array of application scenarios. Video-based social media platforms, in particular, have empowered general users to generate an unprecedented amount of multi-modal data such as texts, audios, images, and the various combinations of them, which enable developers and practitioners to create multi-modal artificial intelligence systems that have already transformed our lives and work, as has been witnessed in the current wave of generative AI applications.

The work in this paper is providing new insights on fusing multi-modal data toward more general tasks beyond sentiment analysis. Our proposed MAVMF algorithm systematically explore the vast feature space that are generated by different modalities and their inherent spatial and temporal relationships. Unfortunately, when compared with other AI tasks such as image recognition and machine translation, the task of fusing multi-modal information for sentiment analysis remains an unsolved problem.

Currently, the underpinning theory regarding how different modalities in text, audio, and visual complement or interfere with each other has not been formalized. However, one future direction on solving the multi-modal fusion challenge could be relying on computation power, similar to what we have observed in large language models (LLM). Without leveraging a larger network with higher computation power may pose limits on the current performance of our proposed method. In another word, we did not consider constraints on algorithm speed given limited computation capacity, as the full MAVMF model has combined features from all proposed modules shown in Figure 2. However, we believe that, as computation is becoming cheaper and more accessible over time, focus in network design and modal fusion should be placed on encompassing all possible interactions among different modalities.

In our next step, we plan to investigate the performance of large foundational models in MSA tasks. Given that the current state-of-the-art performance in existing literature on MSA tasks is still in the 80% range, we will first focus on improving the accuracy and robustness of novel algorithms on MSA tasks. Pre-training with large unlabeled text and image data, foundation models have implicitly encoded large amounts of human knowledge in their weight parameters. It may be promising to adopt them in addressing MSA tasks. Certainly, we can also explore knowledge distillation techniques on successful foundation model-based algorithms to obtain compact apprentice models for more resource constraint scenarios of MSA tasks.

7. Conclusion

Addressing the effective extraction of single-modal features and the efficient integration of multi-modal information, we propose a MSA algorithm MAVMF. First, feature extractors are used to capture single-modal information. Then, for single-modal features, a reduced attention module is used to encode image, while a self-attention module is used

for text. Subsequently, a bidirectional GRU and a fully connected network are applied to extract context-aware discourse features, capturing the context information between discourses in each modality. A sentence-level self-attention module is then used to model different modality information. At the modality level, dual-modality and tri-modality attentions are applied to merge information, and self-attention is used for selecting features with significant contributions in revealing sentiment tendency. Experiments on public datasets prove that, when compared to other deep learning algorithms, the MAVMF model has better or comparable classification performance.

Author Contributions: Conceptualization, Chao He, Lihua Cai and Dingju Zhu; methodology, Chao He, Dongqing Song; software, Chao He and Dongqing Song; validation, Chao He and Dongqing Song; formal analysis, Chao He and Dongqing Song; investigation, Chao He; resources, Lihua Cai, Dingju Zhu, Xinghua Zhang, Yingshan Shen, Chengjie Mao, Huosheng Wen; data curation, Chao He; writing—original draft preparation, He Chao, Dongqing Song, Lihua Cai, Xinghua Zhang, Yingshan Shen, Chengjie Mao, Huosheng Wen; writing—review and editing, Lihua Cai, Dingju Zhu, Xinghua Zhang; visualization, Chao He, Dongqing Song, Xinghua Zhang; supervision, Lihua Cai, Dingju Zhu; project administration, Lihua Cai, Dingju Zhu; funding acquisition, Dingju Zhu. Special thanks to Xinghua Zhang for valuable contributions to the graphical work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Prof. Dingju Zhu.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Both the datasets applied in the current research are open datasets, and can be found on the given link: http://multicomp.cs.cmu.edu/resources/. The datasets are available for download through CMU Multimodal Data SDK GitHub: https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK. The datasets were accessed successfully on 24 January 2024.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169–176.
2. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* **2016**, *31*, 82–88.
3. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* **2017**, *37*, 98–125.
4. Prakash, A.; Chitta, K.; Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7077–7087.
5. Xia, L.; Guanming, L.; Jingjie, Y.; Zhengyan, Z. A Comprehensive Review on Multimodal Dimensional Emotion Prediction. *Acta Automatica Sinica* **2018**, *44*, 2142–2159.
6. Grimaldo, F.; Lozano, M.; Barber, F. MADeM: a multi-modal decision making for social MAS. In Proceedings of the AAMAS (1), 2008, pp. 183–190.
7. Zhu, L.; Zhu, Z.; Zhang, C.; Xu, Y.; Kong, X. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion* **2023**, *95*, 306–325.
8. Xuji, S. A Study on Multimodal Emotion Recognition Based on Text, Speech, and Video. Master's thesis, Shandong University, 2019.
9. Liu, B. *Sentiment analysis and opinion mining*; Springer Nature, 2022.
10. Ting, W.; Wenzhong, Y. A Review of Text Sentiment Analysis Methods. *Journal of Computer Engineering Applications* **2021**, *57*.
11. Jianghao, L.; Yali, G.; Yongmei, Z.; Aimin, Y.; Jin, C. A Study on Constructing an Emotion Dictionary Based on Emoji. *Computer Technology and Development* **2019**, *29*, 181–185.
12. Mike, T.; Kevan, B.; Georgios, P.; Di, C.; Arvid, K. Sentiment in short strength detection informal text. *JASIST* **2010**, *61*, 2544–2558.
13. Saif, H.; He, Y.; Fernandez, M.; Alani, H. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management* **2016**, *52*, 5–19.
14. Yongsui, L.; Liming, W.; Yumei, C.; Zhen, L. A Study on Dynamic Emotion Dictionary Construction Method Based on Bidirectional LSTM. *Microcomputer Systems* **2019**, pp. 503–509.

15. Kanayama, H.; Nasukawa, T. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the Proceedings of the 2006 conference on empirical methods in natural language processing, 2006, pp. 355–363.
16. Rao, Y.; Lei, J.; Wenyin, L.; Li, Q.; Chen, M. Building emotional dictionary for sentiment analysis of online news. *World Wide Web* **2014**, *17*, 723–742.
17. Qi, C.; Li, Z.; Jing, J.; Xinyue, H. A Review Analysis Method Based on Support Vector Machine and Topic Model. *Journal of Software* **2019**, *30*, 1547–1560.
18. Gang, Z.; Zan, X. A Study on Sentiment Analysis Model of Product Reviews Based on Machine Learning. *Research on Information Security* **2017**, *3*, 166–170.
19. Kiritchenko, S.; Zhu, X.; Mohammad, S.M. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* **2014**, *50*, 723–762.
20. Da Silva, N.F.; Hruschka, E.R.; Hruschka Jr, E.R. Tweet sentiment analysis with classifier ensembles. *Decision support systems* **2014**, *66*, 170–179.
21. Kim, Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* **2014**.
22. Okada, M.; Yanagimoto, H.; Hashimoto, K. Sentiment Classification with Gated CNN and Spatial Pyramid Pooling. In Proceedings of the 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, 2018, pp. 133–138.
23. Meng, J.; Long, Y.; Yu, Y.; Zhao, D.; Liu, S. Cross-domain text sentiment analysis based on CNN_FT method. *Information* **2019**, *10*, 162.
24. Jiang, M.; Zhang, W.; Zhang, M.; Wu, J.; Wen, T. An LSTM-CNN attention approach for aspect-level sentiment classification. *Journal of Computational Methods in Sciences and Engineering* **2019**, *19*, 859–868.
25. Hu, Z.; Yue, Y.; Yanyuan, J.; Wenlong, Z. A Study on Sentiment Classification of Online Consumer Reviews Based on Deep LSTM Neural Network. *Chinese Journal of Medical Library and Information* **2018**, *27*, 23–29.
26. Luo, L.x. Network text sentiment analysis method combining LDA text representation and GRU-CNN. *Personal and Ubiquitous Computing* **2019**, *23*, 405–412.
27. Minh, D.L.; Sadeghi-Niaraki, A.; Huy, H.D.; Min, K.; Moon, H. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *Ieee Access* **2018**, *6*, 55392–55404.
28. Zhang, Y.; Jiang, Y.; Tong, Y. Study of sentiment classification for Chinese microblog based on recurrent neural network. *Chinese Journal of Electronics* **2016**, *25*, 601–607.
29. Colombo, C.; Del Bimbo, A.; Pala, P. Semantics in visual information retrieval. *Ieee Multimedia* **1999**, *6*, 38–53.
30. Jindal, S.; Singh, S. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In Proceedings of the 2015 International Conference on Information Processing (ICIP). IEEE, 2015, pp. 447–451.
31. Yang, J.; She, D.; Sun, M.; Cheng, M.M.; Rosin, P.L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia* **2018**, *20*, 2513–2525.
32. Yang, J.; She, D.; Lai, Y.K.; Rosin, P.L.; Yang, M.H. Weakly supervised coupled networks for visual sentiment analysis. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7584–7592.
33. Kumar, A.; Jaiswal, A. Image sentiment analysis using convolutional neural network. In Proceedings of the Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017. Springer, 2018, pp. 464–473.
34. Truong, Q.T.; Lauw, H.W. Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In Proceedings of the Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1274–1282.
35. You, Q.; Jin, H.; Luo, J. Visual sentiment analysis by attending on local image regions. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2017, Vol. 31.
36. Wu, L.; Qi, M.; Jian, M.; Zhang, H. Visual sentiment analysis by combining global and local information. *Neural Processing Letters* **2020**, *51*, 2063–2075.
37. Zheng, R.; Li, W.; Wang, Y. Visual sentiment analysis by leveraging local regions and human faces. In Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26. Springer, 2020, pp. 303–314.
38. Li, L.; Li, S.; Cao, D.; Lin, D. SentiNet: Mining visual sentiment from scratch. In Proceedings of the Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK. Springer, 2017, pp. 309–317.
39. Weifeng, L. A Study on Social Emotion Classification Based on Multimodal Fusion. Master's thesis, Chongqing University of Posts and Telecommunications, 2019.
40. Navas, E.; Hernáez, I.; Luengo, I. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE transactions on audio, speech, and language processing* **2006**, *14*, 1117–1127.
41. Li, B.; Dimitriadis, D.; Stolcke, A. Acoustic and lexical sentiment analysis for customer service calls. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5876–5880.
42. Weiquiqi, L. A Comparative Study of Speech Enhancement Algorithms and Their Applications in Feature Extraction. Master's thesis, Shandong University, 2020.
43. Jun, H.; Yue, L.; Zhongwen, H. Advances in Multimodal Emotion Recognition. *Application Research of Computers/Jisuanji Yingyong Yanjiu* **2018**, *35*.

44. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2539–2544.
45. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* **2016**.
46. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* **2017**.
47. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 163–171.
48. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883.
49. Poria, S.; Peng, H.; Hussain, A.; Howard, N.; Cambria, E. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* **2017**, *261*, 217–230.
50. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
51. Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition* **2023**, *136*, 109259.
52. Yang, B.; Wu, L.; Zhu, J.; Shao, B.; Lin, X.; Liu, T.Y. Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2022**, *30*, 2015–2024.
53. Wu, T.; Peng, J.; Zhang, W.; Zhang, H.; Tan, S.; Yi, F.; Ma, C.; Huang, Y. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems* **2022**, *235*, 107676.
54. Wang, Y.; Li, Y.; Bell, P.; Lai, C. Cross-Attention is Not Enough: Incongruity-Aware Multimodal Sentiment Analysis and Emotion Recognition. *arXiv preprint arXiv:2305.13583* **2023**.
55. He, Y.; Sun, L.; Lian, Z.; Liu, B.; Tao, J.; Wang, M.; Cheng, Y. Multimodal Temporal Attention in Sentiment Analysis. In Proceedings of the Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022, pp. 61–66.
56. Mai, S.; Zeng, Y.; Zheng, S.; Hu, H. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* **2022**.
57. Liu, Q. Study on Emotion Analysis Method Based on Multimodal Information Fusion **2019**.
58. Shenoy, A.; Sardana, A. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *arXiv preprint arXiv:2002.08267* **2020**.
59. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 6818–6825.
60. XueMei, L.; Hong, T.; HongYu, C.; Shanshan, L. Feature Fusion Based on Attention Mechanism - Multi-modal Emotion Analysis Using Bidirectional Gated Recurrent Unit. *Computer Applications* **2021**, *41*, 1268.
61. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
62. Akhtar, M.S.; Chauhan, D.S.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812* **2019**.
63. Yunfeng, S.; Ge, R.; Yong, Y.; Xiaochao, F. Multi-task Multi-modal Emotion Analysis Based on Attention-driven Multilevel Hybrid Fusion. *Application Research of Computers/Jisuanji Yingyong Yanjiu* **2022**, *39*.
64. Xuemei, L.; Hong, T.; Hongyu, C.; Shanshan, L. Attention-based Feature Fusion-Multimodal Sentiment Analysis of Bidirectional Gated Recurrent Units. *Computer Application* **2021**, *41*, 1268.
65. Guangbin, B.; Gangle, L.; Guoxiong, W. Bimodal Interaction Attention for Multi-modal Emotion Analysis. *Computer Science and Exploration* **2022**, *16*, 909.