applied sciences

MDPI

Article

Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement

Ping He *, Huaying Qi, Shiyi Wang and Jiayue Cang

School of Information Technology, Hebei University of Economics and Business, Shijiazhuang 050061, China; huayingqi2023@163.com (H.Q.); wangshiyi_1001@163.com (S.W.); cjyyjc2022@163.com (J.C.)
* Correspondence: pinghe@heuet.edu.cn; Tel.: +86-173-6799-8262

Abstract: Cross-modal sentiment analysis is an emerging research area in natural language processing. The core task of cross-modal fusion lies in cross-modal relationship extraction and joint feature learning. The existing research methods of cross-modal sentiment analysis focus on static text, video, audio, and other modality data but ignore the fact that different modality data are often unaligned in practical applications. There is a long-term time dependence among unaligned data sequences, and it is difficult to explore the interaction between different modalities. The paper proposes a sentiment analysis model (UA-BFET) based on feature enhancement technology in unaligned data scenarios, which can perform sentiment analysis on unaligned text and video modality data in social media. Firstly, the model adds a cyclic memory enhancement network across time steps. Then, the obtained cross-modal fusion features with interaction are applied to the unimodal feature extraction process of the next time step in the Bi-directional Gated Recurrent Unit (Bi-GRU) so that the progressively enhanced unimodal features and cross-modal fusion features continuously complement each other. Secondly, the extracted unimodal text and video features taken jointly from the enhanced cross-modal fusion features are subjected to canonical correlation analysis (CCA) and input into the fully connected layer and Softmax function for sentiment analysis. Through experiments executed on unaligned public datasets MOSI and MOSEI, the UA-BFET model has achieved or even exceeded the sentiment analysis effect of text, video, and audio modality fusion and has outstanding advantages in solving cross-modal sentiment analysis in unaligned data scenarios.

Keywords: cross-modal fusion; gated recurrent unit (GRU) network; multi-head attention; canonical correlation analysis; sentiment analysis

check for updates

Citation: He, P.; Qi, H.; Wang, S.; Cang, J. Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement. *Appl. Sci.* **2023**, *13*, 7489. https://doi.org/10.3390/app13137489

Academic Editor: Vincent A. Cicirello

Received: 22 May 2023
Revised: 21 June 2023
Accepted: 22 June 2023
Published: 25 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the continuous updating and iteration of smart devices, more and more people choose to express their personal views and participate in the discussion of popular social events in the form of texts, images, and short videos on social network platforms, including Weibo, B station, and Douyin. There are a large number of users' sentiment information in these data. The sentiment information is effectively collected and analyzed from these data so that the views, attitudes, and sentiments of users are reflected. Moreover, sentiment information can provide a new research direction for public opinion guidance, service recommendation, and other applications [1–3]. The research of sentiment analysis mainly includes three stages: unimodal, multimodal, and cross-modal. The early stage of unimodal sentiment analysis mainly focuses on analyzing text or image data, which has certain problems, such as insufficient information and semantic ambiguity. Afterward, the sentiment analysis based on text and image mainly falls into the multimodal and cross-modal stage, and the research objects are mostly text data or the data in combination with text and image, which effectively solves the problems of insufficient information and semantic ambiguity in the unimodal sentiment analysis. However, due to the different data sampling rates of different modalities in practical applications, the collected data of different modalities

are frequently unaligned. Although unaligned data sequences have the characteristic of longer length and contain richer information, they also contain a large amount of redundant information. Additionally, there is an asynchronous feature between different modalities, such as the inconsistent appearance of negative words in the text and sad facial expressions in the video, and this asynchronism results in ineffective integration between different modalities. All these problems increase the difficulty of sentiment analysis based on different modalities. Therefore, the motivation of this paper is to obtain more emotional feature representations from unaligned data, to find the correlation and interaction between text and video, and to better achieve emotional analysis.

For the data fusion problem of different modalities in the scenarios of unaligned data sequences, previous studies typically adopted the method of forced word alignment before training the model [4–13]. For the three modalities of text, video, and audio, the pronunciation times of different words in the text are first obtained, and then the video and audio features are forcibly aligned with the pronunciation time intervals of the text. Nevertheless, enforcing word alignment often involves information in relation to the application areas and requires detailed information about the datasets, so it is challenging to implement. Subsequently, Tsai et al. [14] introduced neural machine translation (NMT) as a powerful inspiration and proposed a Multimodal Transformer (MulT) method to fuse the cross-modal information of unaligned data sequences. MulT solves the problem of cross-modal fusion under unaligned data sequences without explicitly aligning the modality data. Fu et al. [15] designed a new lightweight network that introduced a Transformer with added cross-modal blocks to realize complementary learning of different modalities. Notwithstanding, the above studies ignored the importance of unimodal features in the cross-modal fusion of unaligned data sequences.

To extract more concise and multivariate unimodal feature representations, the unaligned data fusion based on feature enhancement technology has been widely studied [16]. He et al. [17] proposed a unimodal enhancement Transformer based on feature enhancement technology, which adds the extracted unimodal features to the unimodal features again to highlight the single peak information and solves the problem of text, video, and audio sentiment analysis under the unaligned data sequence scenarios. Nonetheless, they ignored the interaction between different modalities. Lv et al. [18] proposed a Progressive Modal Enhancement (PMR) method by adding the message centers so that the modal enhancement unit and information updating unit complemented each other. Finally, the enhanced text, video, and audio fusion features are applied to sentiment analysis, and yet the model training complexity of this method increases. Hence, it is still challenging to explore the correlation and interaction between different modalities in the important time steps for cross-modal sentiment analysis in unaligned data sequence scenarios.

To extract effective unimodal feature representations from long sequences, this paper proposes a sentiment analysis model (UA-BFET) based on feature enhancement technology for the cross-modal fusion of text and video in unaligned data scenarios, as shown in Figure 1. Firstly, the model inputs the obtained unimodal vector representation of text and video into a cyclic memory enhancement network across time steps (Cmenats). Secondly, the unimodal fusion features of adjacent time steps are obtained through Bi-GRU, and subsequently, the obtained unimodal fusion features of text and video are utilized as the inputs of the cross-modal hierarchical attention fusion module. Moreover, the obtained cross-modal fusion features with semantic interaction are applied to the unimodal feature extraction process of the next time step in the Bi-GRU system. Thus, the unimodal feature representations and cross-modal fusion feature representations are enhanced. The above operations can reduce the redundant information between the unaligned data sequences and extract effective unimodal feature representations. Moreover, the cross-modal hierarchical attention fusion module, text features, and video features obtained in adjacent time steps interact twice in layers to obtain the correlation and interaction between different modalities on important time steps. The cross-modal hierarchical attention fusion module focuses on the cross-modal interaction in the whole discourse scope, regardless of whether

the different modalities in the unaligned data sequences need to be aligned. In the following, the enhanced cross-modal fusion features combined with the extracted text and video unimodal features are analyzed by canonical correlation analysis (CCA). The loss function of CCA can enhance the correlation between fusion features. Finally, the fusion features with enhanced correlation are input into the fully connected layer and the Softmax function for sentiment analysis.

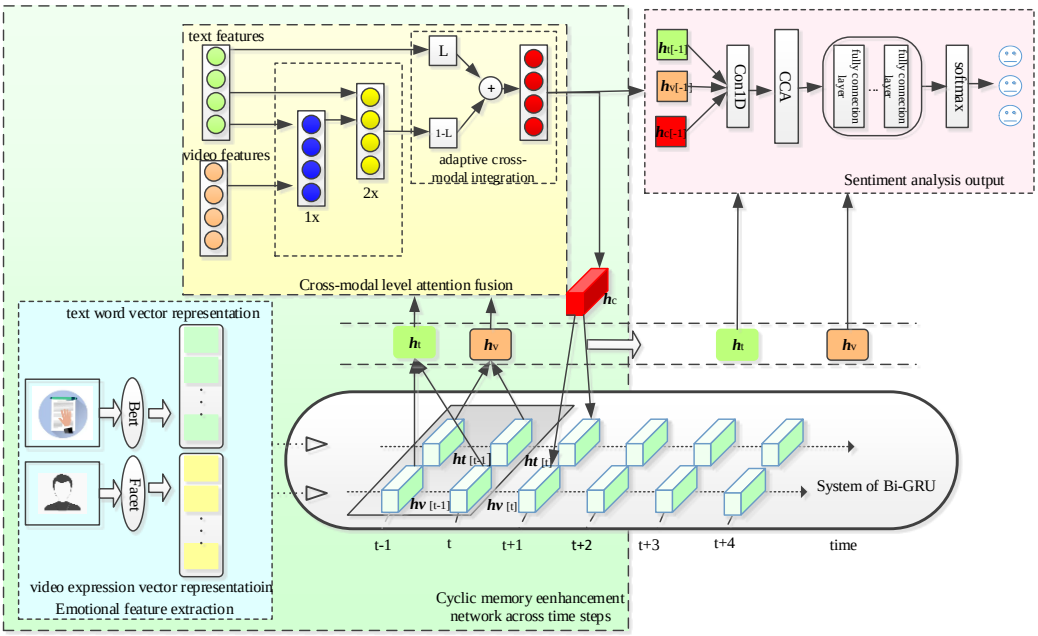


Figure 1. Cross-modal fusion sentiment analysis model of text and video based on feature enhancement technology in unaligned data sequences scenarios.

The main contributions of this paper are summarized as follows:

- This paper proposes a sentiment analysis model (UA-BFET) based on feature enhancement technology in unaligned data scenarios. The model can be used to analyze the sentiment of unaligned text and video modality data in social media.
- It uses a cyclic memory enhancement network across time steps and canonical correlation analysis (CCA) to extract effective unimodal feature representations and explore the correlation and interaction between different modalities.
- The UA-BFET model proposed in this study reaches or even exceeds the sentiment analysis effect of text, video, and audio modalities fusion, which proves the feasibility of the proposed UA-BFET model.

The rest of the paper is organized as follows: Section 2 describes the current related work of sentiment analysis of different modalities. Section 3 gives a description of the proposed UA-BFET sentiment analysis model. Section 4 introduces the experimental designation and verifies the performance of the proposed model. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Multimodal Sentiment Analysis

With the advent of the big data era, the types of data presentation are likewise diversified. For example, text, images, short videos, and other data commonly seen on social media are often presented in pairs or even in combinations of the three. Multimodal sentiment analysis compensates for the problems of insufficient information and semantic ambiguity in unimodal sentiment analysis [19,20]. In 1976, Mcgurk and Macdonald [21] proposed the impact of vision on speech perception, which was used for Audio–Visual

Speech Recognition (AVSR) technology later and became the prototype of the multimodal concept. Since then, multimodal sentiment analysis has undergone a long development. In 2017, Zadeh et al. [22] proposed a new tensor fusion network model for language, visual, and auditory information in online videos, but the model has shortcomings, such as high computational complexity and information redundancy. To solve the above problems, Liu et al. [23] proposed a low-order multimodal fusion method in 2018, which uses low-rank factors to represent the weights among different modalities of text, video, and audio, thus reducing the computational complexity of the tensor fusion model. With the application of the attention mechanism in the area of natural language processing [24], Lin and Meng [25], Guo and Zhang [26], and Fan et al. [27] introduced an attention mechanism to capture text features and image features related to emotions in multimodal sentiment analysis to obtain cross-correlation between different modalities.

Compared with the traditional unimodal sentiment analysis, multimodal sentiment analysis shows better performance, and yet due to the fact that each modality data presents views on things from different perspectives, there is some cross-information between the modalities. However, data redundancy is inevitable when the information intra-modality and the interaction inter-modal are saved. Therefore, cross-modal data fusion is gradually becoming more relevant.

2.2. Cross-Modal Sentiment Analysis

The core task of cross-modal fusion lies in cross-modal relationship extraction and joint feature learning. At the beginning of development, You et al. [28] proposed a cross-modal consistency regression model for text and image sentiment analysis. In 2016, AlphaGO defeated Lee Sedol, and deep learning developed rapidly. Convolution neural networks (CNN) are widely used in cross-modal fusion. In 2016, Cai and Xia [29] used CNN for sentiment analysis of texts and images first, and yet there were also certain problems; for example, text and image could not be directly integrated. To solve the issue of emotional exclusion between text and image, Shen [30] proposed a cross-modal sentiment analysis method based on the fusion of text and image in social media in 2019, and the method uses Continuous Bag of Words (CBOW) and CNN to extract text and image features, respectively. Nevertheless, this method only extracts the high-level semantic features of image sentiment features and ignores the intermediate and low-level features, which also affect the accuracy of sentiment analysis. Therefore, in 2021, Chen et al. [31] used the external convolutional layer of the VGG13 network to obtain the image features of high, low, and medium levels, respectively, and applied them to the cross-modal sentiment analysis of text and images.

Unlike learning different modalities representations from static data (such as text and images), human language contains time series, so sentiment analysis between different modalities requires the fusion of information from time series signals. Previous studies have relied on the assumption that multimodal language sequences are aligned in word parsing and only consider short-term inter-modal interactions. Due to the non-alignment of all modality data in practical applications, it is of more practical significance to study the sentiment analysis of non-alignment data sequence scenarios in case the above assumption does not exist.

2.3. Sentiment Analysis Based on Unaligned Data Sequence Scenarios

The following is an introduction to sentiment analysis based on word alignment and unaligned data sequences.

1. Sentiment analysis based on word alignment: In the scenario of unaligned data sequences, previous studies usually used P2FA [32] to perform the forced word alignment. Previous studies on the fusion of unaligned data sequences have mined the correlation between data of different modalities on the basis of effectively representing unimodal information [33]. Nevertheless, since the performance of sentiment analysis using shallow learning architecture is far from satisfactory, more complex models have been proposed successively [4–13], such as RMFN, MFM, HFFN, RAVEN, etc.

We note that word alignment requires not only detailed information about the domain but also meta-information about the exact time range of words in the datasets, which leads to sentiment analysis not always being feasible in practical applications.

2. Sentiment analysis based on unaligned words: This means that sentiment analysis is performed directly without explicitly aligning sequence data. In 2019, Tsai et al. [14] first extended the NMT Transformer [34] to multimodal sentiment analysis and proposed a multimodal MulT that can learn the interaction information between different modalities and directly perform the sentiment analysis of text, video, and audio without explicitly aligning the sequence data. Although better results are obtained, the complexity of the model increases. Given the above problems, Sahay et al. [35] proposed a low-rank fusion-based Transformer (LMT-MULT) to avoid any excessive parameterization of the model in 2020. Based on this, Fu et al. [15] designed a new lightweight network in 2021, which adopts a Transformer with cross-modal blocks to realize the complementary learning of different modalities. Although the accuracy is not the best, better results are obtained with the least model parameters. However, since the above studies use the same annotation methods in different modalities, the performance in capturing modal differentiation information is poor, and the additional unimodal annotation requires a lot of manpower and time. Based on this, Yu et al. [24] proposed a sentiment analysis method combining self-supervision and multi-task learning in 2021. The difference between different modalities is obtained through independent single-peak supervision, but using fused modal information for sentiment analysis may result in issues such as loss of unimodal information or inclusion of noise in the fused information. To address this problem, in 2021, He et al. [17] proposed a single-peak enhanced Transformer method to carry out sentiment analysis for text, video, and audio. Notwithstanding, the method of feature enhancement based on unimodal is far from enough. In 2022, Lv et al. [18] proposed the progressive modal reinforcement (PMR) method with a message center to enhance the accuracy of sentiment analysis of text, video, and audio through the complementary process of unimodal information and fusion information but ignored the impact of the interaction between different modalities on the final result of sentiment analysis.

Currently, the sentiment analysis in unaligned data sequence scenarios is mostly based on text, video, and audio data, while the above studies (1) ignore the role of considering effective unimodal feature representation and (2) ignore that the correlation and interaction between different modalities in important time steps will lead to certain deviations in the accuracy of sentiment analysis results. Therefore, extracting the heterogeneous relationship between different modality features effectively is still a difficult problem in cross-modal fusion.

3. UA-BFET Sentiment Analysis Model

The paper proposes a cross-modal fusion sentiment analysis model (UA-BFET) of text and video based on feature enhancement technology in unaligned data scenarios. As shown in Figure 1, the model mainly consists of four parts: (1) Unimodal feature extraction. The text and video modality feature tensors extracted from the source data are input into the Bi-GRU for encoding, and the unimodal feature representations with contextual relationships are obtained. (2) Feature enhancement module. This part introduces a cyclic memory enhancement network across time steps. Firstly, text features and video features obtained from adjacent time steps through the Bi-GRU system are input into the cross-modal hierarchical attention fusion module to obtain the cross-modal fusion features with interactions. Secondly, the cross-modal fusion features are applied to the unimodal feature extraction process of the next time step in the Bi-GRU system to solve the problem of unimodal sequence forgetting in the unaligned sequences. The process is repeated so that the incrementally enhanced unimodal features and cross-modal fusion features complement each other continuously to reduce the redundant information between the unaligned sequences and extract an effective unimodal representation related to sentiment.

(3) Cross-modal hierarchical attention fusion. The obtained text features and video features of adjacent time steps are input into the cross-modal hierarchical attention fusion module for hierarchical interaction twice to obtain the cross-modal fusion features that interact with each other on important time steps in the whole unaligned data sequences. It is not necessary to explicitly align different modality data sequences before cross-modal hierarchical attention fusion. (4) Sentiment classification output module. Firstly, the feature enhancement technology based on CCA is used to strengthen the correlation between different modality features at important time steps in long series. Secondly, the enhanced cross-modal fusion features are combined with the extracted unimodal text and video features are subjected to canonical correlation analysis (CCA) and input into the fully connected layer and Softmax function for sentiment analysis.

3.1. Unimodal Feature Extraction

Although the data sequence in the unaligned data scenarios is longer and contains richer information, it is also accompanied by more redundant information. To represent unimodal features effectively, this model mainly analyzes text features and uses video features as the auxiliary for cross-modal sentiment analysis in unaligned data scenarios. As shown in the emotional feature extraction module in Figure 1, given the original text, video modalities (Ti, Vi), for text modality $Ti = (t_{i,1}, \cdots, t_{i,k}, \cdots, t_{i,n}) \in R^{t_1 \times d_1}$, where $t_{i,k}$ represents the kth word of the article ith text. For video modality $Vi = (v_{i,1}, v_{i,2}, \cdots, v_{i,k}, \cdots v_{i,m}) \in R^{t_2 \times d_2}$, where $v_{i,k}$ represents the kth image frame in the ith video fragment. In Ti and Vi, t_1, t_2 represents the length of the text and video modality preprocessing features, respectively, d_1, d_2 represents the dimensions of the text and video modality preprocessing features, respectively, and R represents the set of real numbers.

In this section, we will describe the unimodal feature extraction process, including text feature extraction and video feature extraction.

(1) In the stage of text feature extraction, this model uses *BERT* [19] word embedding to obtain the specific vector representation of text information: $T_{wordi} = BERT(Ti)$, for obtaining the different meanings of the same word in different contexts. Each word is embedded into a d-dimension word vector $t_{wordi,k} \in R^d$, where d represents the feature dimensions of the word vector, so $T_{wordi} = (t_{wordi,1}, \cdots t_{wordi,k}, \cdots t_{wordi,n})$. Then, each word embedding vector is input into the Bi-GRU to obtain a contextual text modality encoding. Both GRU and Long Short Memory (LSTM) networks can solve the problems of gradient vanishing and gradient explosion in traditional Recurrent Neural Networks (RNNs) when dealing with long sequence texts and can better capture dependencies with a larger time step distances in time series. However, the GRU has fewer tensor operations and trains faster in large data volumes [36]. Therefore, this paper selects the GRU deep learning model for text and video modality encoding. The GRU, as a variant of the LSTM, integrates input gates and forgetting gates into update gates and uses hidden states for achieving information transmission. The GRU module structure is shown in Figure 2, which only includes update gates and reset gates.

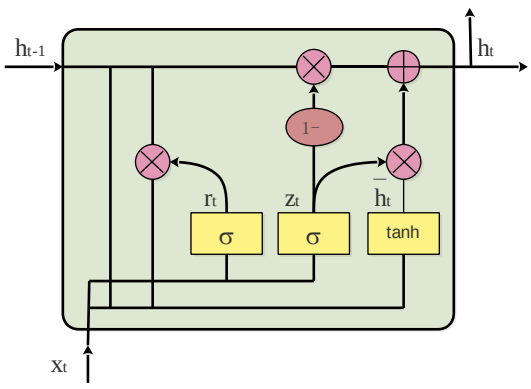


Figure 2. GRU system structure.

Specifically, the update process of the GRU model is shown in Equations (1)–(5). Assume that GRU accepts a word embedding vector $t_{wordi,k}$ as input at time k and outputs a new hidden state vector $h_{i,k}$. And the output at the previous time is denoted as $h_{i,k-1}$.

$$z_{i,k} = \sigma(W_{update} t_{wordi,k} + U_{update} h_{i,k-1}), \tag{1}$$

$$r_{i,k} = \sigma(W_{reset} t_{wordi,k} + U_{reset} h_{i,k-1}), \tag{2}$$

$$\widetilde{h}_{i,k} = \tanh(W t_{wordi,k} + U_{text}(r_{i,k} \times h_{i,k-1})), \tag{3}$$

$$h_{i,k} = (1 - z_{i,k}) \times h_{i,k-1} + z_{i,k} \times \widetilde{h}_{i,k}, \tag{4}$$

$$h_{i,k} = GRU(t_{wordi,k}, h_{i,k-1}) \tag{5}$$

where $z_{i,k}$ is the update gates and denotes whether the previous status needs to be updated; $r_{i,k}$ is the reset gates and indicates whether the previous status needs to be reset. $\widetilde{h}_{i,k}$ represents the hidden state vector to be selected, whose value is determined by $r_{i,k}$. $h_{i,k}$ is the updated hidden state calculated at the time i. W_{update}, U_{update}, W_{reset}, U_{reset}, W, U_{text} are the weight matrix of the corresponding feature, whose parameters are generated by the model training.

Due to the emotion of a word being influenced by the context, the Bi-GRU mechanism is introduced to concatenate the hidden state vectors generated by the forward GRU and backward GRU to obtain the hidden layer output $h_{i,k}^{word}$ of the Bi-GRU, as shown in Equation (6).

$$h_{i,k}^{word} = \overrightarrow{h}_{i.k} \oplus \overleftarrow{h}_{i,k} \tag{6}$$

We use the Bi-GRU end-state hidden vectors as the final word feature representation.

(2) In the stage of video feature extraction, the video is divided into k different segments according to the pauses expressed by characters, denoted as $Vi = (v_{i,1}, v_{i,2}, \cdots, v_{i,k})$, where $v_{i,k}$ represents the *kth* image frame in the *ith* segment. The facial expression features are extracted from the images of different segments by FACET and marked as the initial video modality features, denoted as $V'_{i,k}$. The facial expression features $h_{i,k}^{\text{picture}}$ of each frame are obtained in the video corresponding text words embedding through *Bi-GRU*, as shown in Equation (7).

$$h_{i,k}^{\text{picture}} = Bi - GRU(V'_{i,k}) \tag{7}$$

3.2. Feature Enhancement Module

In order to eliminate the redundant information that existed in the unimodal feature extraction in the unaligned data scenarios and solve the forgetting problem of unimodal sequence, we design a cyclic memory enhancement network across time steps. And to enhance the correlation among the various fusion features, we use the CCA feature enhancement technology to improve the accuracy of the final sentiment analysis. The specific content will be discussed in the following sections.

3.2.1. Cyclic Memory Enhancement Network across Time Steps

In the feature enhancement module, this study designs a cyclic memory enhancement network across time steps. The basic idea is to use the Bi-GRU system to encode unimodal so that the unimodal features can constantly approach the important words over time. At the time step t, $h_m^{[t-1,t]} (m \in \{t, v\})$ denotes the receiving text and video features across two consecutive time steps. First, text features $h_t^{[t-1,t]}$ and video features $h_v^{[t-1,t]}$ across two continuous time steps obtained by the Bi-GRU system are used as the inputs of the cross-modal hierarchical attention module to obtain the gradually enhanced cross-modal fusion

features $h_c \in R^{t_3 \times d_3}$ between text and video features, where d_3 is the feature dimensions of the cross-modal fusion features. Second, the cross-modal fusion features are input into the *Bi-GRU* system at the next time step to update unimodal feature representation. The calculation process is shown in Equation (8):

$$\alpha_{word}^{[t+1]} = Bi - GRU_{update}(h_c; \alpha_{word}^{[1:t]}) \tag{8}$$

where [t + 1] and [1:t] represent time t + 1 from the initial time to time t, respectively. $\alpha_{word}^{[t+1]}$ represents the attention weight of the word obtained from the cross-modal fusion feature h_c at the next time step, t + 1, which is inferred from the attention weight $\alpha_{word}^{[1:t]}$ previously assigned from the initial time to the time t. This denotes that the attention weight of the word at the next time step depends on the influence of the previous cross-modal fusion features on the word. To fully encapsulate these dependencies, the process of attention allocation is executed in a cyclic manner by using the Bi-GRU. And m_t, m_v represent the memory of the Bi-GRU system for text and video modalities, respectively. When t = 0, the initial memory of the Bi-GRU system for text modality is expressed as $h_t^{[0]}$, and the network M is used to map h_t to the Bi-GRU memory space, as shown in Equation (9):

$$h_t^{[0]} = M(h_t; \Theta) \tag{9}$$

The specific intra-modal representation h_t can be dynamically adjusted by the Bi-GRU's memory mechanism.

Since the emotional information and intensity of each word in a sentence are not equal, the model added an additive attention mechanism to highlight the emotional weight of words with different levels of importance in the text. It results that the cross-modal fusion features of different time steps output by the Bi-GRU system are activated by the Softmax function. An attention weight of each word $\alpha_{word}^{[t+1]}$ contained in the sentence is produced, as shown in Equation (11):

$$c_{i,k} = U_{hidden}^T \tan\ h(W_{hidden} h_c^{[1:t]} + b_{hidden}) \tag{10}$$

$$\alpha_{word}^{[t+1]} = \frac{\exp(c_{i,k})}{\sum_{j=0}^{n} \exp(c_{i,k})} \tag{11}$$

where $c_{i,k}$ is the attention value calculated by the model and represents the emotional weight of cross-modal fusion features $h_c^{[1-t]}$ in the sentence from the beginning time to time t. The weight matrix W_{hidden} and bias vector b_{hidden} are used to map $h_c^{[1:t]}$ to the attention space and then multiply the projection with the context vector U_{hidden}.

Next, the importance of the word at time t + 1 in the text is denoted by multiplying the attention weight $\alpha_{word}^{[t+1]}$ with the cascaded intra-modality representation h_t element by element, as shown in Equation (12):

$$\widetilde{h}_t^{[t+1]} = h_{i,k}^{word} \odot \alpha_{word}^{[t+1]} \tag{12}$$

where $\odot$ represents the Hadamard product and $\widetilde{h}_t^{[t+1]}$ denotes the text features participating in the cyclic memory enhancement network across time steps at time t + 1.

After obtaining the text features with weights at time t + 1, it is integrated with the text features h_t cascaded at previous moments. The module is defined by a function f_F, as shown in Equation (13).

$$h_t^{[1:t+1]} = f_F(\widetilde{h}_t^{[t+1]}; h_t^{[1:t]}, \Theta) \tag{13}$$

where $h_t^{[1:t+1]}$ represents the integrated fusion representation of text features from the initial time to time $t+1$. The previous fusion results are integrated by the output gates of the Bi-GRU, so the final vertex outputs of the cyclic memory enhancement network across time steps are obtained, namely the text features $h_t \in R^{t_1 \times d_1'}$, where d_1' is the feature dimensions of the text features. Similarly, the video features of the corresponding text features representation are obtained through the cyclic memory enhancement network cross-time steps and are expressed as $h_v \in R^{t_2 \times d_2'}$, where d_2' is the feature dimensions of the video features.

In the cyclic memory enhancement network across time steps, the gradually enhanced unimodal features and cross-modal fusion features constantly complement each other; that is, h_c enhances h_m, and the enhanced h_m gradually enhances h_c. This can solve the unimodal sequence forgetting problem between the unaligned data sequences and explore the interaction between different modalities on the important time steps so that the cross-modal fusion feature effect is enhanced.

3.2.2. Feature Enhancement Based on Canonical Correlation Analysis

The feature enhancement technology based on CCA is used to enhance the correlation among the fusion features. Specifically, CCA is based on a linear combination of the two random variables. Firstly, the first pair of canonical variables (u_1, v_1) are selected, and a group of correlation coefficients $p(u_1, v_1)$ with the greatest correlation between (u_1, v_1) is obtained. And then, the second pair of canonical variables (u_2, v_2) in this group is selected to make the second pair of canonical variables uncorrelated with the first pair of canonical variables (u_1, v_1). Thus, the second set of correlation coefficients $p(u_2, v_2)$ with the second great correlation is calculated, and so on, until all the correlation coefficients are extracted to simplify the correlations between the original two sets of variables.

As shown in Figure 3, on the basis of $h_t \in R^{t_1 \times d_1'}$, $h_v \in R^{t_2 \times d_2'}$ and $h_c \in R^{t_3 \times d_3}$ obtained in Section 3.2.1, the unimodal feature representations h_t, h_v are first shared by a hard sharing strategy and combined with the obtained cross-modal fusion features h_c. CNN is used to carry out the convolution reduction of the above three features in the time dimension to obtain the one-dimensional global features Vec_h_t, Vec_h_v, Vec_h_c. And then, three 1×1 convolution cores are adopted to convolve and reduce the temporal dimension of three modal features, where the input channels are t_1, t_2, and t_3, respectively, and the output channel is 1, as shown in Equations (14)–(16).

$$Vec_h_t = Conv1D(h_t)^T \in R^{t_1 \times 1} \tag{14}$$

$$Vec_h_v = Conv1D(h_v)^T \in R^{t_2 \times 1} \tag{15}$$

$$Vec_h_c = Conv1D(h_c)^T \in R^{t_3 \times 1} \tag{16}$$

where Vec_h_t, Vec_h_v, and Vec_h_c are three one-dimensional feature vectors.

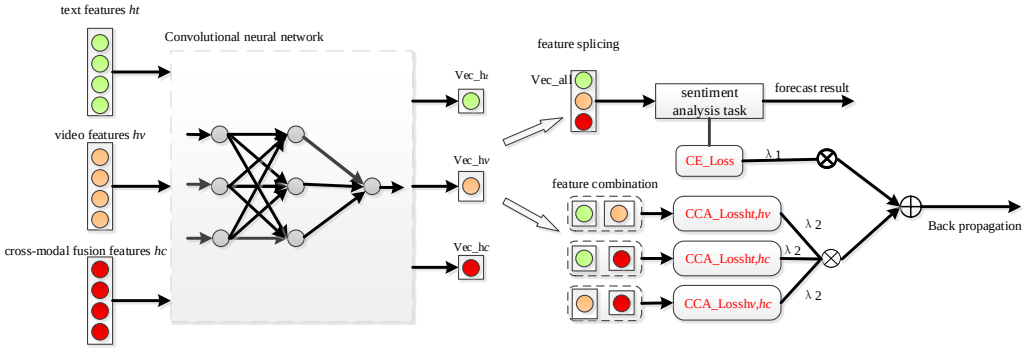


Figure 3. Feature enhancement based on canonical correlation analysis.

Next, the three feature vectors are cascaded into the fully connected layer for sentiment analysis, as shown in Equation (17).

$$Vec_all = Concat(Vec_h_t, Vec_h_v, Vec_h_c) \tag{17}$$

Subsequently, we calculate the canonical correlation coefficients of text features $Vec_h_t \in R^{b_z \times d}$ and video features $Vec_h_v \in R^{b_z \times d}$, where b_z represents the batch size during the training process. After the above two groups of features are given, two groups of neural networks f_1, f_2 are used to carry out nonlinear changes for Vec_h_t, Vec_h_v, respectively, as shown in Equations (18) and (19).

$$Y_{h_t} = f_1(Vec_h_t; \theta_1) \tag{18}$$

$$Y_{h_v} = f_2(Vec_h_v; \theta_2) \tag{19}$$

where θ_1, θ_2 denotes the network parameters of f_1, f_2, respectively.

In the following, the most suitable two groups of network parameters θ_1, θ_2 are found by training two sets of neural networks to make the network output Y_{h_t}, Y_{h_v} in Equations (18) and (19) hold the maximum correlation, as shown in Equation (20).

$$\theta_1^*, \theta_2^* = \arg\max_{\theta_1, \theta_2} \max CCA(Y_{h_t}, Y_{h_v}) \tag{20}$$

Typically, we use the backpropagation algorithm to update the network parameters of the above two networks. Assume M_{11}, M_{22} is the covariance matrix of Y_{h_t}, Y_{h_v}, and M_{12} is the mutual covariance matrix of Y_{h_t}, Y_{h_v}, so the final loss function of CCA is shown in Equation (21).

$$CCA_Loss_{h_t, h_v} = -trace(E^T E)^{\frac{1}{2}} \tag{21}$$

where $E = {M_{11}}^{-\frac{1}{2}} M_{12} {M_{22}}^{\frac{1}{2}}$, so that the above network parameters can update the above results and minimize the loss function.

Similarly, we can obtain another two sets of loss functions $CCA_Loss_{h_t, h_c}$, $CCA_Loss_{h_v, h_c}$ based on CCA.

Specifically, the cyclic memory enhancement network across time steps is used for feature extraction of unimodal and CCA is used for the shared features of modal text features, modal video features, and cross-modal fusion features to solve the unimodal sequence forgetting problem between the unaligned data sequences and enhance the interaction between different modalities obtained at the important time steps. The final goal is to improve the final sentiment analysis effect.

3.3. Cross-Modal Hierarchical Attention Fusion Module

In the cross-modal hierarchical attention fusion module, there is no need to pay attention to the alignment between different modality features. Instead, text features and video features obtained at different time steps directly interact twice hierarchically. Video features and text features were calculated by Multi-head Hierarchical Attention (MHA) to obtain video-level attention distribution under text guidance, which is then recombined with the emotional features of the video to obtain the first fusion results. Similarly, text-level attention distribution under video guidance can still be obtained by MHA, and emotional features after the second fusion can be obtained by combining them with the emotional features of the text. Finally, the final cross-modal fusion features are obtained by adaptive cross-modal integration of text features and text-video features after secondary fusion. The whole process is shown in Figure 4. The redundant information existing in the unaligned data sequence is removed through a cross-modal hierarchical attention fusion module to obtain the interaction between different modalities at important time steps. The module is divided into text-to-video units and video-to-text units. Taking the application of text-to-

video units as an example, the cross-modal hierarchical attention fusion module is divided into three parts: cross-modal interaction, multi-head processing, and adaptive cross-modal integration. The specific process is as follows. The application of video-to-text unit is the same as that of text-to video-unit, and it is no longer described.

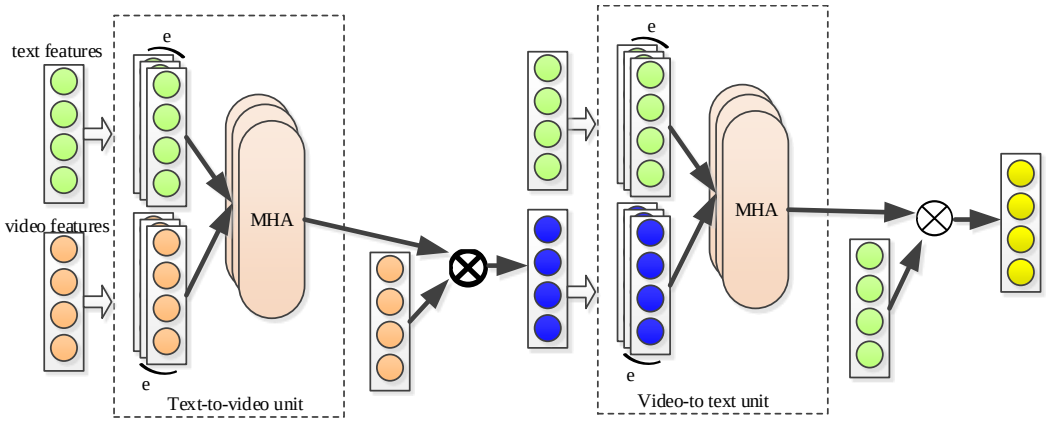


Figure 4. Process of cross-modal hierarchical attention.

3.3.1. Cross-Modal Interaction

In the text-to-video unit, MHA uses 2 times attention on text features and video features to obtain semantic interactions between different modalities. The specific process is shown in Figure 5.

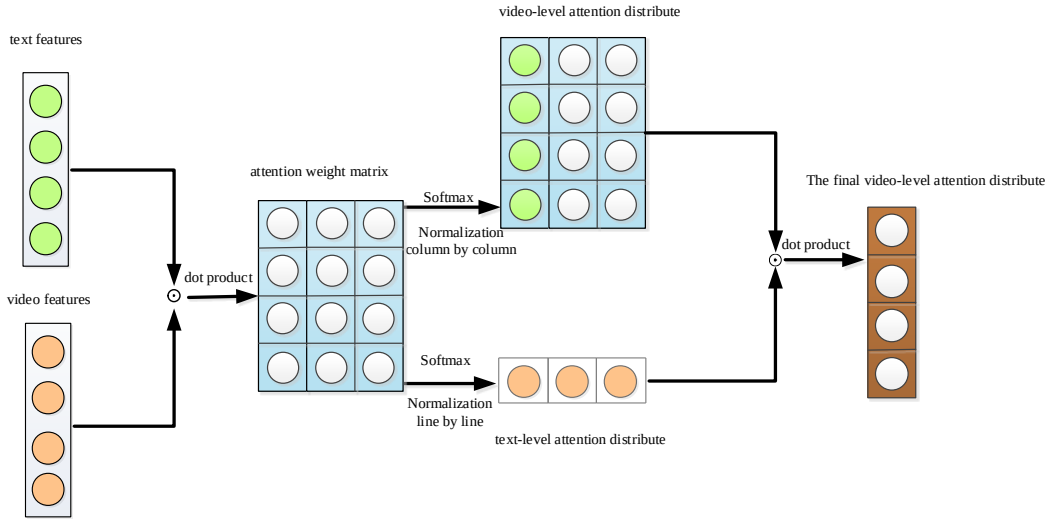


Figure 5. Multi-head Hierarchical Attention, taking the text-to-video unit as an example.

Firstly, the nonlinear projection layer is used to transform the dimension of text features into the same dimension as the video features, as shown in Equation (22), so that the emotional features of the text and video can be confused with hierarchical attention.

$$\hat{h}_t^j = \tanh\left(W_t h_t^j + b_t\right) \tag{22}$$

where W_t and b_t are the weight matrix and deviation, respectively, and $\hat{h}_t^j$ is the representation of the jth text feature h_t^j through the nonlinear projection layer.

Secondly, the attention weight matrix is calculated by the dot product of text and video features, which contains the information of each element, as shown in Equation (23).

$$M(i,j) = \left(h_t^i\right)^T \cdot h_v^j \tag{23}$$

where $M(i,j)$ represents the value calculated by the dot product of the ith feature of text modality and the jth feature of video modality in the attention weight matrix. The larger the value of $M(i,j)$, the stronger the interaction between the two modalities.

Using the Softmax function to normalize the weight matrix column by column, the normalized results of each column are merged to obtain the probability distribution of video-level attention A, as shown in Equation (25). The attention distribution $\alpha(z)$ from text to video is obtained by using the z column as an example, as shown in Equation (24).

$$\alpha(z) = (column)\text{Softmax}(M(1,z), \cdots, M(k,z)) \tag{24}$$

$$A = [\alpha(1), \alpha(2), \cdots, \alpha(n)] \tag{25}$$

As with video-level attention, the probability distribution of text-level attention β is obtained by using the Softmax function to normalize the weight matrix line by line, as shown in Equation (26). To obtain more information with semantic interaction, the average values are calculated by $\beta(z)$ to obtain average text-level attention β, as shown in Equation (27).

$$\beta(z) = (row)\text{Softmax}(M(z,1), \cdots, M(z,n)) \tag{26}$$

$$\beta = \frac{1}{\mathrm{n}} \sum_{z=1}^{k} \beta(z) \tag{27}$$

Finally, to obtain the weight distribution of the influence of important emotional words in the sentence on video features, the final video-level attention s is computed by the dot product of A and β, as shown in Equation (28).

$$s = A^T \cdot \beta \tag{28}$$

3.3.2. Multi-Head Processing

In this model, multi-head attention is used to enhance the fusion effect of text features and video features in order to obtain key unimodal features to enhance the effect of sentiment analysis, as shown in Figure 6.

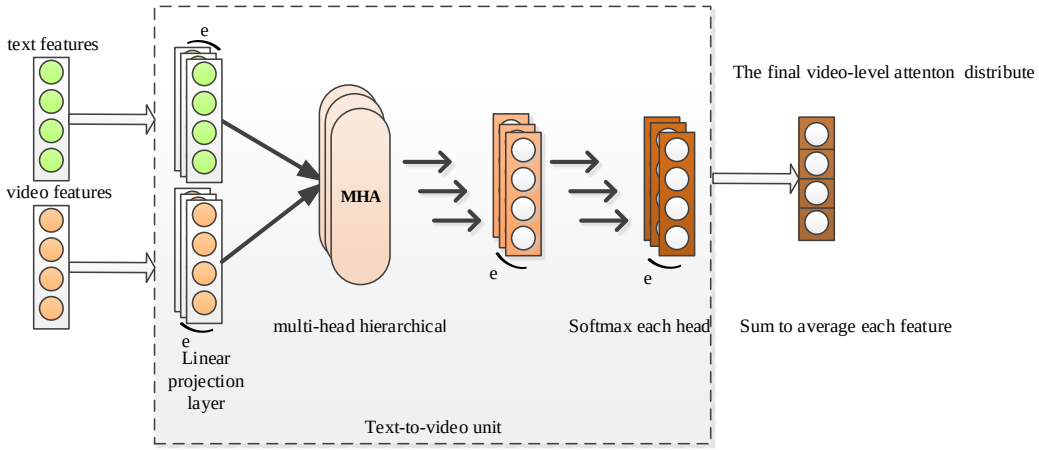


Figure 6. Multi-head processing, using the text-to-video unit as an example.

Firstly, text features and video features are input into the linear projection layer e times. The goal is to convert the dimension of text features into the same dimension as video features and obtain more key information in different representation subspaces. Moreover,

the normalized values of each header calculated by the Softmax function line by line are summed and averaged to obtain video-level attention *Svid*, as shown in Equation (29).

$$Svid = MultiHead(h_v, h_t) = \frac{1}{n}\sum_{i=1}^{|e|} \text{softmax}(head_i) \tag{29}$$

$$head_i = Atten(h_v W_i^{h_v}, h_t W_i^{h_{t1}}) \tag{30}$$

where $W_i^{h_v}$ and $W_i^{h_t}$ represents the trainable weight of the network.

Secondly, video-level attention *Svid* and video features h_v are combined again. And the first fusion result h_g is obtained by the product function of elements, as shown in Equation (31).

$$h_g = Svid \odot h_v \tag{31}$$

where $\odot$ represents the product of elements.

Similarly, in the video-to-text unit, the first fusion result h_g and text features h_t are input into MHA to obtain the final text-level attention *Stex* and then combined with text features to obtain the secondary fusion features h_f, as shown in Equations (32)–(34).

$$Stex = MultiHead(h_t, h_g) = \frac{1}{n}\sum_{i=1}^{|e|} softmax(head_i) \tag{32}$$

$$head_i = Atten(h_t W_i^{h_t}, h_g W_i^{h_{tg}}) \tag{33}$$

$$h_f = Stex \odot h_t \tag{34}$$

In this section, cross-modal interaction information is obtained by hierarchical interaction twice. Among them, multi-head processing can obtain more key emotional information.

3.3.3. Adaptive Cross-Modal Integration

To suppress the problem of carrying irrelevant noise during fusion, text features h_t and cross-modal features after the secondary fusion h_f are merged again by using the method of self-adjusting output ratio [37], and the final cross-modal fusion features h_c are obtained, as shown in Equation (36). According to the existing studies [30,38], when the text features h_t and the features h_f after the secondary fusion do not match at all in content, the scalar of the text is 1, and the scalar of the video is 0. The integral function is defined as follows:

$$\gamma = \sigma\left(W_\gamma h_t + U_\gamma h_f + b_\gamma\right) \tag{35}$$

$$h_c = \gamma \odot h_t + (1 - \gamma) \odot h_f \tag{36}$$

where W_γ, U_γ is an adaptive weight matrix, b_γ is a bias, the range of fused values is (0, 1), and σ is the element-level sigmoid function.

3.4. Sentiment Classification Output Module

In the sentiment classification output module, text features h_t, video features h_v, and cross-modal fusion features h_c are input into the fully connected layer and Softmax function for sentiment analysis. Since the sentiment scores have positive and negative sentiment polarity, $score \in [-N, N]$, and N is a positive integer. Thus, the Relu function is chosen as the activation function of the fully connected layer, as shown in Equations (37) and (38).

$$Vec_{all} = concat(h_t[-1], h_v[-1], h_c[-1]) \tag{37}$$

$$Score = W_{o_2}(\mathrm{Re}lu(Vec_{all}W_{o_1} + b_1)) + b_2 \tag{38}$$

where $[-1]$ indicates that we specify the last frame of the three groups' features as the training target to participate in network training, W_{o_1}, W_{o_2}, b_1, and b_2 represent the training weight and network parameters, respectively.

Secondly, the loss function of sentiment analysis is trained together with the loss function of CCA. Depending on the different weight values of the loss function, it constantly strengthens the correlation between various modality features when the network focuses on the training tasks of sentiment analysis. Considering the evaluation criteria, cross-entropy is used as a loss function to train the model, as shown in Equation (39).

$$CE_Loss = \frac{1}{D}\sum_{k=1}^{d} y'_{i,k}\log(y_{i,k}) \tag{39}$$

where D is the total number of training samples, d is the number of sentiment categories, $y_{i,k}$ is the probability number of the ith text–video pair belonging to the kth category, and $y'_{i,k}$ is the label value. And the positive class value is 1; otherwise, it is 0.

Finally, combined with Formulas (21) and (39), the complete loss function is as follows:

$$Loss = \lambda_1 CE_loss + \lambda_2(CCA_loss_{h_t,h_v} + CCA_loss_{h_t,h_c} + CCA_loss_{h_v,h_c}) \tag{40}$$

where λ_1, λ_2 represent the weights of the MSE loss function and CCA loss function, respectively.

In the process of network training, after weighting according to Formula (40), the whole loss function is used for the gradient descent method to update the network parameters.

4. Experimental Evaluation

4.1. Datasets

Because CMU-MOSI and CMU-MOSEI datasets are widely used in sentiment analysis, this study selects CMU-MOSI and CMU-MOSEI datasets to verify the feasibility of the proposed model. The designed experiments are verified from the perspective of word aligned and unaligned.

CMU-MOSI: As a data set widely used in the area of sentiment analysis, CMU-MOSI is edited from 93 videos posted by 89 different narrators on YouTube, including a total of 2199 discourse videos. Among them, 1284 discourse video segments are used as training samples, 229 discourse video segments are used as verification samples, and 686 discourse video segments are used as test samples. Video features are extracted with a sampling rate of 15 Hz, while text modality is segmented by word and represented as discrete word embedding.

CMU-MOSEI: As a data set also used in the area of sentiment analysis, CMU-MOSEI is much larger than CMU-MOSI in terms of data volume. CMU-MOSEI is edited from 3837 videos posted by 1000 different narrators on YouTube, including a total of 22,856 discourse videos. In the training process of the model, 16,326 discourse video segments are selected as training samples, 1871 discourse video segments are selected as verification samples, and 4659 discourse video segments are selected as test samples. Video features are extracted with a sampling rate of 15 Hz. The datasets are shown in Table 1.

Table 1. Data set statistics.

Classification	CMU-MOSI	CMU-MOSEI
Train set	1284	16,326
Valid set	229	1871
Test set	686	4659
Data set summation	2199	22,856

4.2. Experimental Setting

To achieve better classification results in the CMU-MOSI and CMU-MOSEI datasets, this experiment has carried out some hyperparameter settings, as shown in Table 2. The hyperparameters are determined in the valid set. Batch size represents the number of training samples in each batch, Learning_rate_bert represents the learning rate of BERT word embedding, Epochs represents the number of training iterations, and hidden sizes of the Bi-GRU network layer are used to transmit input information. The kernel size is used to process input sequences of text, video, and cross-modal fusion features before the fully connected layer. The model is trained under Pytorch1.10.2 architecture. For a fair comparison, the average performance after five runs is selected as the final result in model training. The experiment adopts the Dropout method to prevent overfitting.

Table 2. The hyperparameter settings adopted in the CMU-MOSI and CMU-MOSEI datasets.

Setting	CMU-MOSI	CMU-MOSEI
Optimizer	Adam	Adam
Batch size	32	32
Learning_rate_bert	5×10^{-5}	5×10^{-5}
Epochs	100	120
Bi-GRU hidden sizes	32	32
Attention head	8	10
Kernel size ($h_t/h_v/h_c$)	1/1/1	1/1/1
Dropout	0.1	0.1

Under the CMU-MOSI and CMU-MOSEI datasets, the experiment optimizer was Adam, and the Batch size value was set to 32. After running the experiment five times, we set Epochs to 100 and 120, respectively. The hidden sizes of the Bi-GRU network layer are set to 16, 32, and 64, respectively, during the experiment process. When the value is 32, the classification effect is the best. Hence, the hidden sizes of the Bi-GRU network layer are set to 32. The value of the Learning_rate_bert parameter was determined by a similar experimental method. The kernel size is 1×1 convolution kernel, and one-dimensional global vectors of text features, video features, and cross-modal fusion features are obtained and sent into the fully connected layer. The value analysis of parameter Dropout is shown in Section 4.4.2.

4.3. Evaluation Metrics

When evaluating the performance of the model, the evaluation index proposed by the studies of previous unaligned sentiment analysis [9,39–41] is selected to illustrate the experimental results of this study from two perspectives of classification and regression.

As for the classification evaluation index, Accuracy, Recall, and F1-Score are used to evaluate the performance of sentiment analysis. F1-Score is an index that comprehensively considers both Precision and Recall rate, as shown in Equations (41)–(44).

$$P_A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FN} + N_{FP} + N_{TN}} \tag{41}$$

$$P_R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{42}$$

$$P_P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{43}$$

$$P_F = \frac{2 \times P_R \times P_P}{P_R + P_P} \tag{44}$$

where N_{TP} is the number of samples correctly labeled as a positive example, N_{FP} is the number of samples incorrectly labeled as a positive example when they are actually a

negative example, N_{TN} is the number of samples correctly labeled as a negative example, and N_{FN} is the number of samples incorrectly labeled as a negative example when they are actually a positive example. P_A stands for Accuracy, P_R stands for Recall rate, P_P stands for Precision, and P_F stands for F1-Score.

For the regression evaluation indicators, the mean absolute error (MAE) and Pearson correlation (Corr) are used to evaluate the performance of sentiment analysis. With the exception of MAE, higher values indicate better performance for all indicators.

4.4. Quantitative Analysis

The proposed UA-BFET model is compared with some existing state-of-the-art baseline models based on unaligned data sequences. The methods based on word forced alignment include early fusion LSTM (EF-LSTM), late fusion (LF-LSTM), Multimodal Factorization Model (MFM) [8], Recurrent Attended Variation Embedding Network (RAVEN) [13], Hierarchical Feature Fusion Network (HFFN) [10], Multimodal Cyclic Translation Network model (MCTN) [9], Modality-Invariant and -Specific Representations (MISA) [11], and Multimodal Split Attention Fusion (MSAF) [42]. The methods based on word unaligned include Multimodal Transformer (MulT) [14], Low-Rank Matrix Factorization Transformer (LMT-MULT) [35], Learn Modality-Fused Representations with CB-Transformer (LMR-CBT) [15], Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) [24], Unimodal Reinforced Transformer (UR-Transformer) [17], Progressive Modal Enhancement (PMR) [18], and Weighted Cross-Modal Attention Mechanism [36].

4.4.1. Word Alignment Setting

The word alignment settings require additional steps to manually align visual streams in word resolution and perform a cross-modal fusion of text and video on the time step of word alignment. In this study, the experimental results of each method in two datasets of MOSI and MOSEI are shown in Tables 3 and 4, respectively. Compared with other baselines, the Accuracy and F1-Score of the UA-BFET model are only lower than those of the MSAF model with word alignment settings.

Table 3. Cross-modal sentiment analysis results in aligned/unaligned MOSI data sets. The bold part represents the optimal values for all baseline models.

Model	Acc-2/%	F1-Score/%	MAE	Corr
CMU-MOSI (Word alignment)				
EF-LSTM	75.3	75.2	1.023	0.608
LF-LSTM	76.8	76.7	1.015	0.625
MFM [8]	78.1	78.1	0.951	0.662
RAVEN [13]	78.0	76.6	0.915	0.691
HFFN [10]	80.2	80.3	-	-
MCTN [9]	79.3	79.1	0.909	0.676
MISA [11]	81.8	81.7	0.783	0.761
MSAF [42]	-	-	-	-
UA-BFET (ours)	**84.7**	**84.7**	**0.721**	**0.797**
CMU-MOSI (unaligned)				
MulT [14]	81.1	81.0	0.889	0.686
LMT-MulT [35]	78.5	78.5	0.957	0.681
LMR-CBT [15]	81.2	81.0	-	-
Self-MM [24]	84.0	84.4	0.713	**0.798**
UR-Transformer [17]	82.2	82.4	**0.603**	0.662
PMR [18]	82.4	82.1	-	-
Weighted Cross-modal Attention Echanism [36]	84.7	84.6	0.716	0.786
UA-BFET (ours)	**84.7**	**84.7**	0.721	0.797

Table 4. Cross-modal sentiment analysis results in aligned/unaligned MOSEI data sets. The bold part represents the optimal values for all baseline models.

Model	Acc-2/%	F1-Score/%	MAE	Corr
CMU-MOSEI (Word alignment)				
EF-LSTM	78.2	77.9	0.642	0.616
LF-LSTM	80.6	80.6	0.619	0.659
MFM [8]	-	-	0.568	0.717
RAVEN [13]	79.1	79.5	0.614	0.662
HFFN [10]	60.4	59.1	-	-
MCTN [9]	79.8	80.6	0.609	0.670
MISA [11]	83.6	83.8	0.555	0.756
MSAF [42]	**85.5**	**85.5**	0.559	0.738
UA-BFET (ours)	84.5	83.9	**0.424**	**0.793**
CMU-MOSEI (unaligned)				
MulT [14]	81.6	81.6	0.591	0.694
LMT-MulT [35]	80.8	81.3	0.620	0.668
LMR-CBT [15]	80.9	81.5	-	-
Self-MM [24]	82.8	82.5	0.530	0.765
UR-Transformer [17]	81.8	81.8	0.597	0.671
PMR [18]	83.1	82.8	-	-
Weighted Cross-modal Attention Echanism [36]	83.8	83.8	0.547	0.751
UA-BFET (ours)	**84.5**	**83.9**	**0.424**	**0.793**

4.4.2. Unaligned Settings

The unaligned settings do not require explicit alignment of different modality data sequences. The experimental results are shown in Tables 3 and 4. Compared with other baseline models in unaligned settings, the UA-BFET model has the best performance under all four different evaluation metrics.

Firstly, compared with the MISA model with optimal performance based on word alignment settings in the CMU-MOSI dataset, the proposed UA-BFET model improves the Acc-2 by 2.9% and the F1-Score by 3.0%. In the CMU-MOSEI dataset, based on the MSAF model with the best performance in word alignment settings, the UA-BFET model has decreased by 1% in Acc-2 and 1.6% in F1-Score. And it has the best performance in MAE and Corr. The experimental results show that the UA-BFET model, which adds the cyclic memory enhancement network across time steps and the feature enhancement technique of CCA, can solve the forgetting problem of unimodal sequence and hold better interaction between different modalities at important time steps. This is consistent with the research motivation of this paper.

Secondly, based on the unaligned settings, the accuracy of the UA-BFET model is consistent with the overall best-performing Weighted Cross-Modal Attention Mechanism model in the CMU-MOSI dataset, with a 0.1% improvement in F1-Score and a 1.1% improvement in Corr evaluation index. As shown in Tables 3 and 4, it can be seen from the experimental results that the UA-BFET model can achieve the same or even better results than the fusion of three different modalities. In the CMU-MOSEI dataset, compared with the optimal Weighted Cross-Modal Attention Mechanism model, the Acc-2 of the UA-BFET model increased by 0.7%, and the F1-Score increased by 0.1%. In the experiment, we compared the experimental results of the UA-BFET model in the unaligned data scenarios with other word alignment and unaligned baseline models, respectively. It can be proved that the model performance improvement of the UA-BFET model is small compared with the baseline model in the unaligned setting, and the model performance improvement is significant when compared with the baseline model in the word alignment setting. The reason is that the number of parameters and computational complexity of the cross-modal sentiment analysis model are larger in the unaligned setting than in the word alignment setting.

In addition, the experiment uses the Dropout method to prevent overfitting. The value of the parameter Dropout affects the final output of the model. On the one hand, the larger the value of Dropout, the less overfitting the model will hold, and yet the model's generalization ability will also decrease. Because Dropout randomly discards some neurons, resulting in the loss of some important emotional information. On the other hand, if the value of Dropout is too small, the model may be overfitting. Hence, in the experiment, when the value of Dropout is 0.1, each evaluation index realizes the optimal value, and the generalization ability of the model can be improved without overfitting. Additionally, we note that the proposed method is more effective in obtaining critical information about unimodal sequences when the modal sequences involved are too long.

4.5. Ablation Results

4.5.1. The Validity of CCA

In the sentiment classification output module, this study uses the CCA enhancement technology to maximize the correlation between text modality features, video modality features, and cross-modal fusion features. As shown in the upper part of Table 5, compared with the model without CCA, the accuracy of the UA-BFET model with CCA increased by 0.6% in Acc-2 and 0.8% in F1-Score. This indicates that using CCA to enhance the correlation between different modalities can effectively improve the accuracy of cross-modal sentiment analysis in unaligned data scenarios.

Table 5. Ablation results (taking MOSI data set as an example).

Input Modal	Acc-2/%	F1-Score/%	MAE	Corr
Without CCA	84.1	83.9	0.413	0.794
With CCA	84.7	84.7	0.424	0.793
text	82.8	82.8	0.708	0.800
video	82.9	75.2	0.702	0.801
audio	83.2	75.5	0.697	0.811
audio + video	83.8	83.8	0.716	0.801
text + video	84.7	84.7	0.424	0.793
text + audio	84.5	84.5	0.727	0.793

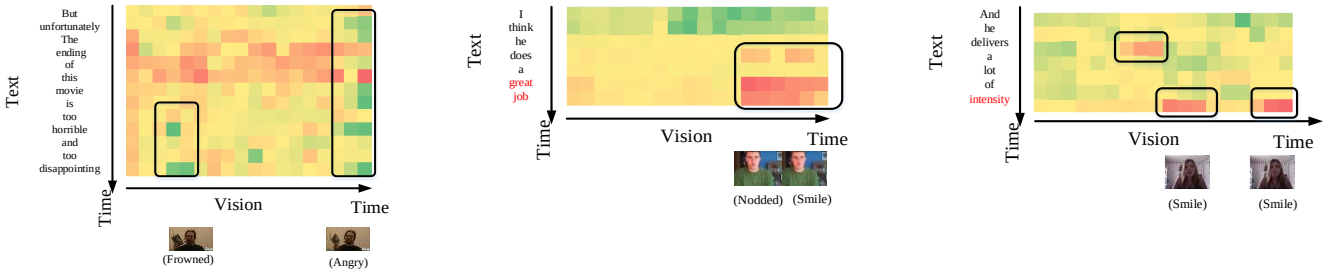
4.5.2. The Validity of Cross-Modal Fusion of Text and Video in Unaligned Data Scenarios

To further verify the effectiveness of sentiment analysis of text and video cross-modal fusion, some experiments were set on three unimodal and pairwise dual-modality data, and the average of multiple experimental results are taken as the final results. As shown in the bottom half of Table 5, it can be seen that the effect of the pairwise dual-modalities sentiment analysis is obviously superior to the unimodal sentiment analysis. Additionally, the overall performance of the combination of text and video or text and audio is better than that of audio and video. Considering that text features are extracted after a lot of training, while audio and video features are extracted manually. Therefore, text modality has more influence on the results of sentiment analysis than video and audio modality. Based on the development status of social platforms and the performance of the combination of text and video in the experiment, it is proved that it is feasible to choose text and video modality as the input of the UA-BFET model in this paper.

4.6. Case Study

To further illustrate the effect of the UA-BFET model on cross-modal sentiment analysis in the unaligned data scenarios, we visualized the attention weight of interaction between text and video cross-modal elements in the dataset. As shown in Figure 7, in Case a, when the words "unfortunately", "horrible", and "disappointing" expressed negative emotions, they were also accompanied by a series of facial expressions, such as "frowned" and "angry" in the video modality, so the model could accurately predict the negative emotional attitude.

In Case b, the overall polarity of the video modality is biased towards negative emotion due to the drooping eyebrows presented by the people in the video. Nevertheless, the polarity of cross-modal fusion features is consistent with the polarity of the true emotion. It can be seen that the video modality is canceled out in the final cross-modal fusion. It is shown that the model can explore the correlation and interaction between text representing emotions and corresponding video segments in long sequences. In Case c, the text modality "deliver a lot of intensity" exhibits ambiguous emotions, and the smiling facial expression removes ambiguities that appear in the text modality, as shown by the section circled in the black block diagram in Case c, which exhibits positive emotions overall. These further indicate that the model can enhance the unimodal features and effectively improve the cross-modal sentiment analysis effect in the unaligned data scenarios by using the correlation and interaction between different modalities.



(**a**) (**b**) (**c**)

Figure 7. Attention weight visualization of interactions between text and video cross-modal elements in CMU-MOSI and CMU-MOSEI data sets. It can be seen that in the unaligned scenarios, the UA-BFET model proposed in this paper can realize the association of emotive text words with facial expression changes in video modalities. In the three cases, the three-color scale of red, yellow, and green is used to represent the emotional polarity of specific cases; that is, the color tending to red corresponds to positive emotion, while the color tending to green corresponds to negative emotion. The darker the color, the greater the emotional value. The part circled by the black block diagram represents the final emotional representation of the interaction between text modality and video modality. In case (**a**), it can be seen the part circled by the black block diagram is biased towards green and the overall is biased towards negative emotion, while the overall color of case (**b**,**c**) is biased towards red and the overall is biased towards positive emotion.

Through experiments executing on unaligned public datasets MOSI and MOSEI, the UA-BFET model has obtained better sentiment analysis results in unaligned data scenarios, which proves the advantages of sentiment analysis from the perspective of extracting effective unimodal features and interaction between different modalities. Notwithstanding, when sarcasm, comparison, and irregular text appear in the text modality, the model cannot always accurately predict the sentiment, so it can be considered to introduce an external knowledge base to solve the above problems. We believe that this study is an important step forward in solving cross-modal sentiment analysis in unaligned data scenarios.

5. Conclusions and Future Work

This paper proposes a sentiment analysis model (UA-BFET) based on feature enhancement technology for the cross-modal fusion of text and video in unaligned data scenarios. Specifically, the forgetting problem of unimodal sequences in unaligned sequences is solved based on the cyclic memory enhancement network across time steps, and the unimodal loss is reduced. The enhancement technology based on CCA strengthens the correlation between different modalities at the important time steps and makes the obtained modal features tend to be true emotions. Experimental results on unaligned MOSI and MOSEI datasets show that the proposed UA-BFET model can enhance the weight of emotion-related unimodal features in unaligned data sequences and provide effective help for the

subsequent acquisition of correlation and interaction between different modalities, as well as the accuracy of cross-modal sentiment analysis.

We believe that the UA-BFET model for cross-modal sentiment analysis in unaligned data scenarios provides more possibilities for specific areas such as cross-modal visual Q&A, image and text retrieval, and recommendation systems. Next, we will consider introducing more external knowledge to better handle sentiment analysis of irregular texts, satirical phrases, and comparative sentences in real-world corpora.

Author Contributions: Conceptualization, P.H. and H.Q.; methodology, P.H.; software, H.Q.; validation, S.W., J.C. and H.Q.; formal analysis, P.H.; investigation, H.Q.; resources, H.Q.; data curation, H.Q.; writing—original draft preparation, H.Q. and S.W.; writing—review and editing, P.H. and J.C.; visualization, H.Q.; supervision, P.H.; project administration, P.H.; funding acquisition, P.H. All authors have read and agreed to the published version of the manuscript.

Funding: Scientific Research and Development Program Project of the Hebei University of Economics and Business (2022ZD09).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data can be provided by the authors upon request.

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

References

1. Chen, L.; Guan, Z.Y.; He, J.H.; Peng, J.Y. A survey on sentiment classification. *J. Comput. Res. Dev.* **2017**, *54*, 1150–1170.
2. Lehrer, S.F.; Xie, T. The bigger picture: Combining econometrics with analytics improve forecasts of movie success. *Manag. Sci.* **2022**, *68*, 189–210. [CrossRef]
3. O'Connor, B.; Balasubramanyan, R.; Routledgeb, B.R.; Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010. [CrossRef]
4. 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 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017. [CrossRef]
5. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018. [CrossRef]
6. Gu, Y.; Yang, K.N.; Fu, S.Y.; Chen, S.H.; Li, X.Y.; Marsic, I. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
7. Liang, P.P.; Liu, Z.Y.; Zadeh, A.; Morency, L.-P. Multimodal Language Analysis with Recurrent Multistage Fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018. [CrossRef]
8. Tsai, Y.-H.H.; Liang, P.P.; Zadeh, A.; Morency, L.-P.; Salakhutdinov, R. Learning factorized multimodal representations. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019. [CrossRef]
9. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.-P.; Poczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019. [CrossRef]
10. Mai, S.J.; Hu, H.F.; Xing, S.L. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019. [CrossRef]
11. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, Seattle, WA, USA, 12–16 October 2022. [CrossRef]
12. Dai, W.L.; Cahyawijaya, S.; Bang, Y.J.; Fung, P. Weakly-supervised Multi-task Learning for Multimodal Affect Recognition. *arXiv* **2021**, arXiv:2104.11560. [CrossRef]
13. Wang, Y.S.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019. [CrossRef]

14. Tsai, Y.-H.H.; Bai, S.J.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019. [CrossRef]
15. Fu, Z.W.; Liu, F.; Wang, H.Y.; Shen, S.Y.; Zhang, J.H.; Qi, J.Y.; Fu, X.L.; Zhou, A.M. LMR-CBT: Learning Modality-fused Representations with CB-Transformer for Multimodal Emotion Recognition from Unaligned Multimodal Sequences. *arXiv* **2021**, arXiv:2112.01697. [CrossRef]
16. Zhang, R.; Xue, C.G.; Qi, Q.F.; Lin, L.Y.; Zhang, J.; Zhang, L. Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 1915. [CrossRef]
17. He, J.X.; Mai, S.J.; Hu, H.F. A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis. *IEEE Signal Process. Lett.* **2021**, *28*, 992–996. [CrossRef]
18. Lv, F.M.; Chen, X.; Huang, Y.Y.; Duan, L.X.; Lin, G.S. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021. [CrossRef]
19. Chen, J.Y.; Yan, S.K.; Wong, K.-C. Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis. *Neural Comput. Appl.* **2020**, *32*, 10809–10818. [CrossRef]
20. Rao, T.R.; Li, X.X.; Xu, M. Learning multi-level deep representations for image emotion classification. *Neural Process. Lett.* **2020**, *51*, 2043–2061. [CrossRef]
21. Mcgurk, H.; Macdonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef]
22. Zadeh, A.; Chen, M.H.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017. [CrossRef]
23. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018. [CrossRef]
24. Yu, W.M.; Xu, H.; Yuan, Z.Q.; Wu, J.L. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021. [CrossRef]
25. Lin, M.H.; Meng, Z.Q. Multimodal sentiment analysis based on attention neural network. *Comput. Sci.* **2020**, *47*, 508–514, 548.
26. Guo, K.X.; Zhang, Y.X. Visual-textual sentiment analysis method based on multi-level spatial attention. *J. Comput. Appl.* **2021**, *41*, 2835–2841.
27. Fan, T.; Wu, P.; Wang, H.; Ling, C. Sentiment analysis of online users based on multimodal co-attention. *J. China Soc. Sci. Tech. Inf.* **2021**, *40*, 656–665.
28. You, Q.Z.; Luo, J.B.; Jin, H.L.; Yang, J.C. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the AAAI conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015. [CrossRef]
29. Cai, G.Y.; Xia, B.B. Multimedia sentiment analysis based on convolutional neural network. *J. Comput. Appl.* **2016**, *36*, 428–431, 477.
30. Shen, Z.Q. A cross-modal social media sentiment analysis method based on the fusion of image and text. *Softw. Guide* **2019**, *18*, 9–13, 16.
31. Chen, Q.H.; Sun, J.J.; Sun, L.; Jia, Y.B. Image-text sentiment analysis based on multi-layer cross-modal attention fusion. *J. Zhejiang Sci-Tech Univ. (Nat. Sci. Ed.)* **2022**, *47*, 85–94.
32. Yuan, J.H.; Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **2008**, *123*, 3878. [CrossRef]
33. Zeng, Z.H.; Tu, J.L.; Pianfetti, B.; Liu, M.; Zhang, T.; Zhang, Z.Q.; Huang, T.S.; Levinson, S.E. Audio-visual affect recognition through multi-stream fused HMM for HCI. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005. [CrossRef]
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems, Long Beach, CA, USA, 4–9 December 2017. [CrossRef]
35. Sahay, S.; Okur, E.; Kumar, S.H.; Nachman, L. Low Rank Fusion based Transformers for Multimodal Sequences. *arXiv* **2020**, arXiv:2007.02038. [CrossRef]
36. Chen, Q.P.; Huang, G.M.; Wang, Y.B. The Weighted Cross-Modal Attention Mechanism With Sentiment Prediction Auxiliary Task for Multimodal Sentiment Analysis. *IEEE ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2689–2695. [CrossRef]
37. Tian, Y.; Sun, X.; Yu, H.F.; Li, Y.; Fu, K. Hierarchical self-adaptation network for multimodal named entity recognition in social media. *Neurocomputing* **2021**, *439*, 12–21. [CrossRef]
38. Yang, Q.; Zhang, Y.W.; Zhu, L.; Wu, T. Text sentiment analysis based on fusion of attention mechnism and BiGRU. *Comput. Sci.* **2021**, *48*, 307–311.
39. Guo, X.B.; Kong, W.-K.A.; Kot, A.C. Deep Multimodal Sequence Fusion by Regularized Expressive Representation Distillation. *IEEE Trans. Multimed.* 2022. [CrossRef]
40. Han, W.; Chen, H.; Poria, S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Virtual Event / Punta Cana, Dominican Republic, 7–11 November 2021. [CrossRef]

41. Zhao, Z.P.; Wang, K. Unaligned Multimodal Sequences for Depression Assessment From Speech. In Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, UK, 11–15 July 2022. [CrossRef]
42. Su, L.; Hu, C.Q.; Li, G.F.; Cao, D.P. MSAF: Multimodal split attention fusion. *arXiv* **2020**, arXiv:2012.07175. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.