electronics

MDPI

Article

Text-Centric Multimodal Contrastive Learning for Sentiment Analysis

Heng Peng [1], Xue Gu [2], Jian Li [1], Zhaodan Wang [3] and Hao Xu [1,*]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China; pengheng21@mails.jlu.edu.cn (H.P.); lijian20@mails.jlu.edu.cn (J.L.)
[2] Department of Industrial Electronics, School of Engineering, University of Minho, 4800-058 Guimares, Portugal; id9267@alunos.uminho.pt
[3] College of Aviation Foundation, Aviation University of Air Force, Changchun 130012, China; zdwang99@sina.com
* Correspondence: xuhao@jlu.edu.cn; Tel.: +86-431-8516-6358

Abstract: Multimodal sentiment analysis aims to acquire and integrate sentimental cues from different modalities to identify the sentiment expressed in multimodal data. Despite the widespread adoption of pre-trained language models in recent years to enhance model performance, current research in multimodal sentiment analysis still faces several challenges. Firstly, although pre-trained language models have significantly elevated the density and quality of text features, the present models adhere to a balanced design strategy that lacks a concentrated focus on textual content. Secondly, prevalent feature fusion methods often hinge on spatial consistency assumptions, neglecting essential information about modality interactions and sample relationships within the feature space. In order to surmount these challenges, we propose a text-centric multimodal contrastive learning framework (TCMCL). This framework centers around text and augments text features separately from audio and visual perspectives. In order to effectively learn feature space information from different cross-modal augmented text features, we devised two contrastive learning tasks based on instance prediction and sentiment polarity; this promotes implicit multimodal fusion and obtains more abstract and stable sentiment representations. Our model demonstrates performance that surpasses the current state-of-the-art methods on both the CMU-MOSI and CMU-MOSEI datasets.

Keywords: multimodal sentiment analysis; contrastive learning; pre-trained language model; feature fusion

Citation: Peng, H.; Gu, X.; Li, J.; Wang, Z.; Xu, H. Text-Centric Multimodal Contrastive Learning for Sentiment Analysis. *Electronics* **2024**, *13*, 1149. https://doi.org/10.3390/electronics13061149

Academic Editor: Daniele Riboni

Received: 21 February 2024
Revised: 14 March 2024
Accepted: 19 March 2024
Published: 21 March 2024

1. Introduction

As a critical technology in human–computer interactions, sentiment analysis is widely applied in fields such as disease treatment, public opinion analysis, and fatigue monitoring [1]. In its early stages, this work primarily relied on single modalities, such as text [2–4] or audio [5–7]. In recent years, with the development of multimedia and short video platforms, multimodal sentiment data have exploded, presenting new challenges and attracting numerous scholars to engage in multimodal sentiment analysis. Unlike uni-modal sentiment recognition, multimodal sentiment recognition enhances sentimental feature understanding by introducing information from other modalities, providing a more comprehensive and nuanced sentimental understanding. This approach addresses information gaps and ambiguity issues in uni-modal sentiment analysis [8].

Early research in this field focused on extracting and mining sentimental information, often neglecting attention to multimodal fusion perception [9–12]. Subsequently, with the rise of deep learning, multimodal sentiment analysis began leveraging deep neural networks to extract and fuse features from different modalities automatically. The emphasis shifted toward designing efficient neural network architectures to integrate various types of data, resulting in the proposal of several excellent multimodal fusion methods [13–17].

With the advent of pre-trained language models, such as the BERT series [18–20], the models trained on large corpora demonstrated the efficient capture of textual context representations and superior performance in natural language processing tasks. This significant progress led many researchers to combine pre-trained language models with deep learning networks, leveraging both technologies' strengths for augmented feature extraction and fusion. Some studies [21–26] view pre-trained language models as excellent alternatives to traditional word embedding methods such as GloVe [27] and Word2Vec [28]. In contrast, others design models with fine-tuned pre-trained language models as the core framework [29–31].

Although prior research has leveraged pre-trained language models to enhance multimodal sentiment analysis (MSA) tasks, the design of these models has not yet broken the paradigm of modal balance. Researchers overlook the fact that the density and quality of textual information far surpass those of the other two modalities (audio and visual) after the improvement provided by pre-trained language models. This viewpoint is further corroborated by extensive ablation experiments in the realm of MSA [23,32]. Therefore, we adopt a text-centric model design philosophy, focusing on the textual modality and utilizing audio and visual modalities as auxiliary components to augment textual information for sentiment analysis.

Moreover, current research on multimodal feature fusion often relies on the assumption of feature space consistency. Directly processing different modal features under this assumption overlooks the distribution differences that carry implicit inter-sample relationships and modality interaction information. In this paper, we use contrastive learning to capture these differences and apply them to sentiment representation learning. Many studies have demonstrated the superiority of contrastive learning in information acquisition within feature spaces [33–35].

Based on the considerations above, we propose a text-centric multimodal contrastive learning (TCMCL) framework for multimodal sentiment analysis. The framework is centered around text and incorporates a cross-modal text augmentation module based on Siamese network architecture, which augments textual features separately with audio and visual conditions to expand nontextual sentiment clues. Within this architecture, we introduce two contrastive learning tasks to delve deeper into sample correlations and modality interaction information. Following the idea of text-centric approaches, the instance prediction-based contrastive learning (IPCL) task achieves the implicit fusion and alignment of multimodal sentiment information through the cross-prediction of the two augmented text features. Supervised by label information, sentiment polarity-based contrastive learning (SPCL) efficiently learns subtle feature differences among samples with different sentiment polarities, maintaining the model's sensitivity to sentiment expression. Our experimental results demonstrate the superiority of our framework in multimodal sentiment analysis tasks.

In summary, our contributions are the following:

- We have introduced a text-centric multimodal contrastive learning (TCMCL) framework for sentiment analysis. The framework uses audio and visual modal information to provide auxiliary augmentations to textual content.
- We have proposed two contrastive learning strategies based on instance prediction and sentiment polarity, aiming to unearth deep sentimental space information and achieve implicit cross-modal fusion alignment.
- Our model achieves state-of-the-art performance on the CMU-MOSI and CMU-MOSEI datasets.

The rest of the paper is structured as follows: Section 2 provides an overview of the existing research on multimodal sentiment analysis and comparative learning, and Section 3 presents a detailed explanation of our proposed TCMCL framework. Section 4 presents our experimental results and provides thorough analyses. Finally, Section 5 summarizes the findings and conclusions of this paper.

2. Related Work

In this section, we primarily discuss the work related to the TCMCL framework, including topics related to multimodal sentiment analysis and contrastive learning.

2.1. Multimodal Sentiment Analysis

As an evolving research domain, multimodal sentiment analysis integrates various modalities such as text, visuals, and audio to understand and analyze sentimental expressions comprehensively. Early research often employed simple concatenation or cascading strategies for data fusion [9–12].

However, with the rapid development of deep learning technologies, the methodology for multimodal sentiment analysis has also been continuously innovating. For instance, methods based on long short-term memory (LSTM) [36], gated recurrent unit (GRU) [37], and convolutional neural networks (CNNs) effectively perform feature extraction and the fusion of multimodal data through carefully designed combinations. Poria et al. [38] successfully utilized deep networks based on RNNs for multimodal sentiment recognition. Zadeh et al. [14] employed a tensor fusion network (TFN) for end-to-end learning, achieving a more detailed and comprehensive understanding of sentiments through the outer product fusion of tensor representations from multiple perceptual modalities. Zadeh et al. [16] developed the multi-memory fusion network (MFN) model by utilizing memory attention networks and multiview gated memories for temporal information modeling in multi-perspective temporal learning. Kumar et al. [39] achieved deep multimodal feature vector fusion by introducing learnable gating mechanisms, self-attended context representations, and recurrent layer-based self and gated cross-fusion. Paraskevopoulos et al. [17] proposed a neural architecture for multimodal fusion, utilizing a feedback mechanism in the forward pass during network training to capture top-down cross-modal interactions. Subsequently, numerous studies have employed even more novel approaches. For example, Nguyen et al. [40] combined 3D convolutional neural networks (C3Ds) and deep belief networks (DBNs), introducing deep spatiotemporal features and effectively fusing visual and audio feature vectors through bi-linear pooling techniques. Zhang et al. [41] proposed quantum-inspired multimodal networks (QMNs) by utilizing the mathematical formalism of quantum theory (QT) to simulate interactions between modalities and speakers.

The rise of pre-trained language models, such as ELMo [42], GPT [43], and BERT [18], has brought about significant transformations in the field of natural language processing (NLP). These models are pre-trained on large corpora and capture complex semantic information and contextual relationships, substantially improving text analysis performance. They have not only propelled the development of NLP but have also been widely applied in research on multimodal sentiment analysis. Many studies replace traditional word embedding methods, such as GloVe and Word2Vec, with pre-trained language models for text feature extraction. For instance, researchers such as Han, Yu, and Hazarika [21–23] have incorporated BERT as the feature extraction tool for textual information in their model designs. Zeng et al. [44] pre-trained BERT on MSA corpora, resulting in improved performance in text feature extraction. Simultaneously, other studies adopt fine-tuned pre-trained language models as the main framework for cross-modal fusion. For instance, Yang et al. [29] used an improved BERT model—ALBERT—and fused text and audio information through model fine-tuning and masked multimodal attention mechanisms. Rahman et al. [30] introduced a multimodal adaptive gate (MAG), enabling BERT and XLNet to accept and process multimodal nonlinguistic data during the fine-tuning phase. Kim et al. [31] proposed All-modalities-in-One BERT for multimodal sentiment analysis, a model pre-trained on two tasks, multimodal masked language modeling (MMLM) and cross-modal alignment prediction (AP), aiming to capture dependencies between modalities. These studies consistently indicate that pre-trained language models have significant potential to enhance text representations, with textual information deserving more attention compared with audio and video modalities. In order to build upon this observation, our research steadfastly adopts a text-centric approach. Contrary to the balanced model

design strategy, our framework is primarily centered around text, utilizing audio and visual modalities to augment the expressive capability of textual sentiment features, and is ultimately employed for sentiment analysis.

2.2. Contrastive Learning

As a self-supervised learning approach, contrastive learning aims to learn more meaningful and robust data representations by exploring the similarity or dissimilarity between data samples. Self-supervised contrastive learning has been extensively studied in the field of computer vision (CV), leading to a series of models based on its fundamental principles. The InstDisc model [45] introduced pivotal concepts such as individual discrimination tasks and a memory bank, which have since served as fundamental elements in self-supervised contrast learning. The individual discrimination task partitions each instance (e.g., an image) into distinct categories, enabling the model to discern individual instances. The positive samples are the instances themselves (after data augmentation), and the negative samples comprise other instances in the dataset. Moreover, InstDisc employs a memory bank to archive features for all instances, integrating proximal regularization constraints to facilitate momentum-based feature updates within the memory bank, thus alleviating any storage pressures on the model. In diverging from InstDisc, the MoCo model [33] replaces the traditional memory bank with queues as an auxiliary data structure for negative sample storage. It also employs momentum encoders in lieu of conventional constraint terms, enabling encoder updates instead of feature modifications. The SimCLR model [34] eschews the use of a memory bank, opting instead for a larger batch size and generating pairs of positive and negative samples within the same batch of data. Its minimalist model architecture consists solely of a pair of shared encoders and a projection layer. Subsequent models, such as BYOL [46] and SimSiam [35], eschew negative samples entirely, adopting a twin network architecture that utilizes only positive samples for training. As self-supervised representation learning methods, they have surpassed traditional supervised learning methods in multiple visual tasks.

Moreover, natural language processing (NLP) has spawned a wealth of good work, but its model structure is still similar to the several architectures mentioned above. The key distinction in the application of self-supervised contrastive learning in these two fields lies in the data augmentation methods. In image processing, data augmentation predominantly involves cropping, rotating, Gaussian blurring, and adding Gaussian noise, whereas these operations are not directly applicable to text. The ConSERT model [47] utilizes adversarial attack, token shuffling, cutoff (token cutoff and feature cutoff), and dropout to augment text, achieving favorable results through contrastive learning. Conversely, SimCSE [48] employs a more straightforward and elegant approach, utilizing two different dropout augmentations on textual data to construct pairs of positive samples. These studies demonstrate the effectiveness of contrast learning in extracting high-level abstract features and provide new directions for downstream research. Inspired by the above work, we applied the principles of self-supervised contrast learning to the multimodal sentiment analysis framework and designed a cross-modal text augmentation method. In addition, we extended the application of contrast learning from self-supervised learning to supervised learning, using sentiment labeling to enhance the adaptability of features to domain tasks.

3. Methods

In this section, we provide an overarching introduction to the proposed text-centric multimodal contrastive learning (TCMCL) framework for sentiment analysis. Subsequently, we present detailed information on the feature extraction process for three modalities and describe the network structure of the cross-modal text augmentation module. Additionally, we introduce two contrastive learning tasks conducted in this module. Finally, we explain the calculation of the overall training loss.

3.1. Overall Architecture

Figure 1 provides a comprehensive overview of the TCMCL framework, which is centered around text and comprises two main components: the feature extraction module and the cross-modal text augmentation mechanism (highlighted in purple). In the feature extraction module, we utilized a pre-trained BERT model for text feature extraction, COVAREP [49] for audio feature extraction, and FACET [50] for visual feature extraction. Subsequently, we performed word-level alignment and length segmentation on the three types of features. The cross-modal text augmentation module employs a Siamese network architecture, where each branch consists of a text augmentation encoder and a projection layer. The branches share weights and perform text augmentation encoding from both audio and visual aspects. Additionally, we designed two contrastive learning tasks based on instance prediction and sentiment polarity to achieve implicit modal alignment and learn deeper representations of sentiment features. Finally, we merged the augmented text features from both branches to form the ultimate multimodal representation and used this for sentiment analysis computation.

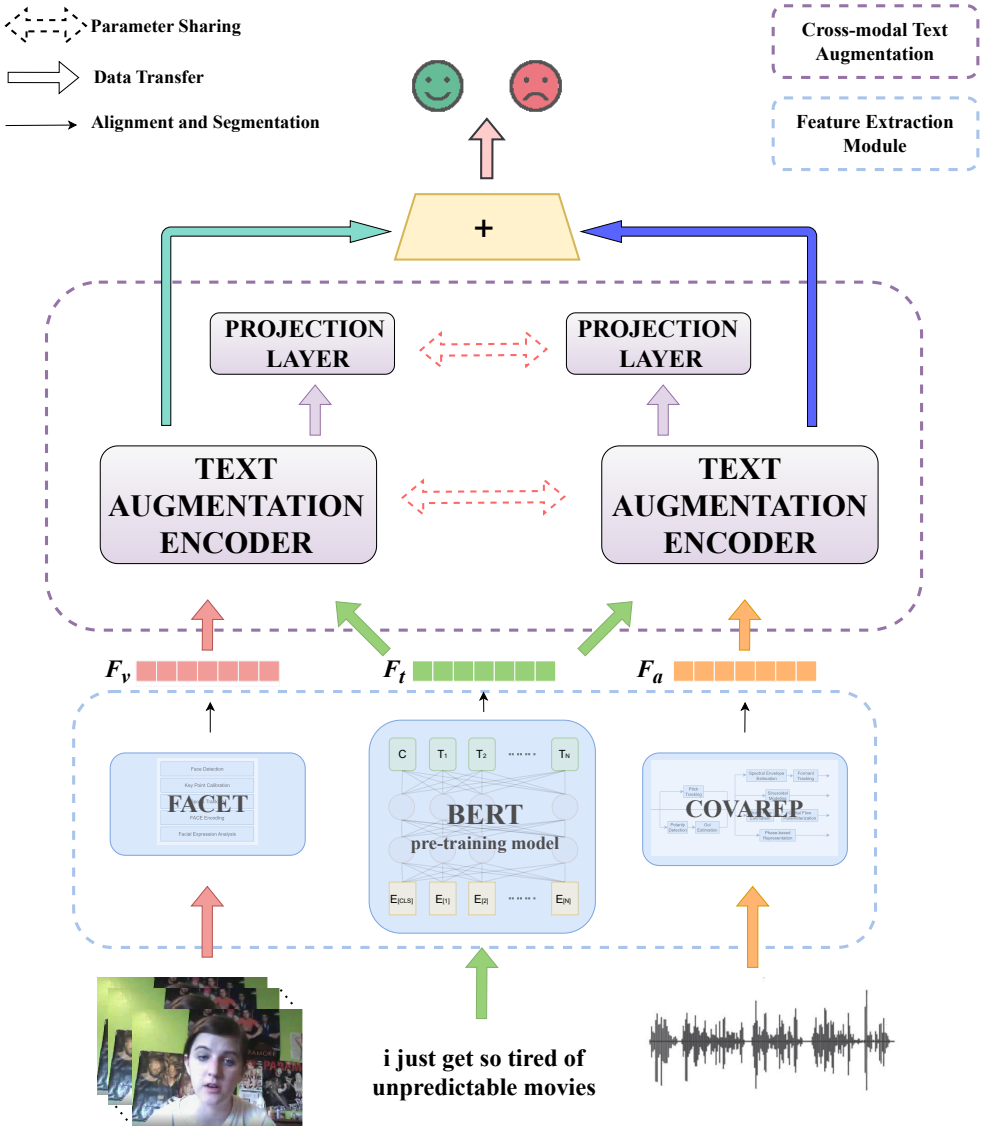


Figure 1. A diagram of the overall architecture of TCMCL.

3.2. Single-Modal Feature Extraction

To begin with, in order to effectively extract features from different modalities, we adopted distinct approaches. In the textual domain, we employed the pre-trained BERT model to process the text, obtaining embeddings and contextual information for each word in the sentence. Audio features, on the other hand, are more abundant and complex. In order to acquire comprehensive audio information, we opted for the COVAREP audio processing toolkit to extract the following features: the fundamental frequency, quasi-open quotient, normalized amplitude quotient, glottal source parameters (H1H2, Rd, and Rd conf), voice/unvoiced (VUV), multidimensional quality (MDQ), first three formants, pitch strength parameters (PSP), harmonics-to-noise ratio (HMPDM 0–24), and the harmonics-to-noise ratio difference (HM-PDD 0–12), as well as the spectral tilt/slope of wavelet responses (peak/slope) and mel-frequency cepstral coefficients (MCEP 0–24). Feature extraction for visual information is equally crucial in multimodal analysis. In this study, we employed the FACET library for the batch processing of visual information, extracting critical visual features such as facial action units, landmarks, head pose, gaze tracking, and histograms of oriented gradient (HOG) features.

Upon obtaining feature sequences from the three modalities of text, audio, and visual information, we aligned all three modalities following the convention in [51]. This word-level alignment ensures synchronization between the audio and visual features and each word in the text, enabling the model to establish a word-level feature sequence for subsequent computation and analysis. Furthermore, to optimize the model's processing capability, reduce the computational burden, and eliminate redundant information, we standardized the length of the feature sequences for all modalities to a fixed scale (N). This standardization was implemented through segmentation and discarding data beyond the sequence length N, effectively eliminating some blank feature vectors.

The above operations can be simply expressed as the following:

$$F_t = f_{\text{aas}}(BERT(T)) \in \mathbb{R}^{N \times d_t} \tag{1}$$

$$F_a = f_{\text{aas}}(COVAREP(A)) \in \mathbb{R}^{N \times d_a} \tag{2}$$

$$F_v = f_{\text{aas}}(FACET(V)) \in \mathbb{R}^{N \times d_v} \tag{3}$$

where T, A, and V represent the raw data of the text, audio, and visual modalities, respectively. The function f_{aas} symbolizes word-level alignment and segmentation operations. Ultimately, we have derived three modal key sequence features: F_t, F_a, and F_v. The sequence length of these features is denoted as N, whereas d_t, d_a, and d_v represent the feature dimensions of the three modalities.

The extracted feature information was fed into the subsequent stages of the TCMCL framework, offering substantial support for our multimodal sentiment analysis task.

3.3. Cross-Modal Text Augmentation

In order to further extract sentiment-related commonalities across the three modalities and capture the feature space differences arising from modality interactions, we designed a cross-modal text augmentation module. This module adopts a Siamese network architecture comprising two components: the text augmentation encoder and the projection layer. Specifically, the text augmentation encoder treats the features from auxiliary modalities (audio and visual) as supplements to and augmentations of text features, thus integrating the features from auxiliary modalities into text features to enrich them while preserving textual contextual information. The projection layer performs spatial mapping on the augmented text features to acquire high-dimensional abstract information. Within this module, we combine two contrastive learning tasks based on instance prediction and sentiment polarity, achieving the implicit multimodal fusion of sentiment information and augmenting the model's ability to differentiate between different sentimental states, thereby improving overall performance.

3.3.1. Siamese Network Structure

The cross-modal text augmentation module consists mainly of a text augmentation encoder and a projection layer. Due to parameter sharing between branches, we illustrate the network structure and parameter-passing mechanism by using the text and visual modality branches as examples, as depicted in Figure 2.

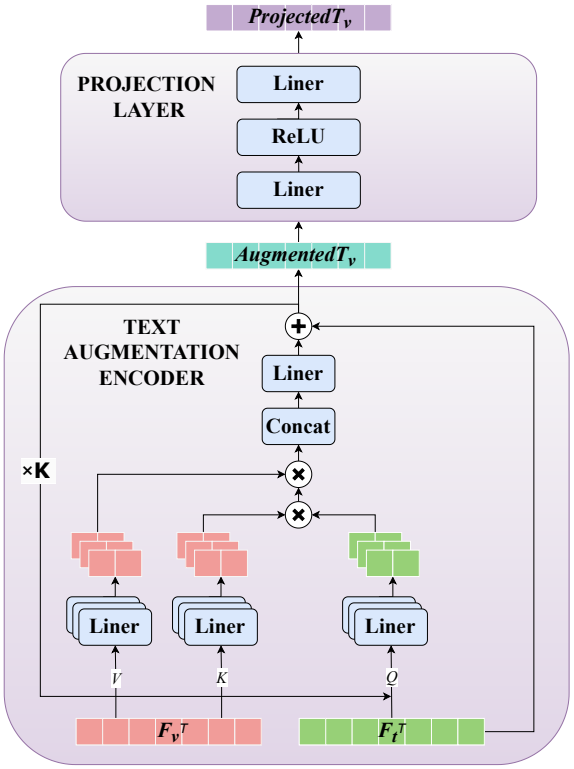


Figure 2. Illustration of the cross-modal text augmentation module's text–visual branch.

Initially, we outlined the structure of the text augmentation encoder, a pivotal component in enhancing the multimodal sentiment analysis procedure. As illustrated in Figure 2, the text augmentation encoder comprises K stacked multi-head cross-attention units. Cross-attention mechanisms are commonly employed to extract modality interaction information and perform feature fusion. In the TCMCL model, these mechanisms are designed to delve into the augmentation effects of audio and visual information on text. Consequently, to preserve the contextual representation advantages of text features better than traditional methods that emphasize dependencies between sequences, our attention mechanisms focus more on the information correlation between different modality features. In the preceding feature extraction stage, we obtained diverse and rich information from the audio and visual modalities using specialized tools. By leveraging feature-level attention mechanisms, we can more clearly identify which features of the auxiliary modalities are most advantageous for text representation, effectively augmenting text features.

In order to emphasize attention toward the features, we first transposed the text and visual features to obtain $F_a{}^\top$ and $F_v{}^\top$. Subsequently, we utilized $F_a{}^\top$ as Q, $F_v{}^\top$ as K, and V for multi-head attention computation. Such transformation facilitates feature attention during the subsequent attention computation and changes the dimensions of the different modality features from their respective d to the sequence length, N, due to transposition. This enables direct parameter sharing between the two branches of the text augmentation encoder. The specific computation process of the text augmentation encoder is as follows.

Initially, we map Q, K, and V into distinct subspaces through linear transformations:

$$\begin{aligned} Q_i &= QW_i^Q \\ K_i &= KW_i^K \\ V_i &= VW_i^V \end{aligned} \tag{4}$$

where W_i^Q, W_i^K, and W_i^V represent the learnable weight matrix, and i denotes the i-th attention head. After linear mapping, attention is computed for each head:

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{5}$$

At this stage, $Q_i K_i^T$ captures feature-level attention when moving from text to visuals, with a matrix size of $d_t \times d_a$, instead of the sequence-level attention of $N \times N$. Here, the denominator is employed to scale the dot product results, preventing gradient vanishing issues. The softmax function is applied to each row to transform scores into probabilities.

Subsequently, the outputs from all heads are concatenated and passed through another linear transformation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{6}$$

where $\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$, W^O represents another learnable weight matrix, and h denotes the total number of heads. Thus, we obtain the feature-level attention when moving from the text to the visual modality $MultiHead(Q, K, V)$.

Finally, we generate residual connection between Q and $MultiHead(Q, K, V)$ to maximize the preservation of textual features' characteristics, forming the output of the current multi-head cross-attention unit:

$$\text{AugmentedT}_v = Q + \text{MultiHead}(Q, K, V) \tag{7}$$

AugmentedT_v will serve as the input Q for the next iteration of the cross-attention unit, as illustrated in Figure 2. After iterating K times, we obtain our feature fusion module's final output: AugmentedT_v.

Through a feature-level attention mechanism, we successfully integrate and complement the targeted augmentation of the visual-assisted modality's feature information into text information. By maintaining the predominant role of textual information, this fusion strategy introduces beneficial attention to feature-level visual information. We consider the module output as an augmented encoding of the text, which, in conjunction with the augmented encoding from another branch, elevates the sentimental expression of the text, laying a solid foundation for downstream tasks.

Next, we devised a projection layer to map the aforementioned augmented features into a new space, aiming to capture abstract and robust sentiment features better. The specific operations of the projection layer are expressed as follows:

$$\begin{aligned} \text{Z}_1 &= \text{Linear}(\text{AugmentedT}_v) \\ \text{ReLU}_1 &= \text{ReLU}(\text{Z}_1) \\ \text{ProjectedT}_v &= \text{Linear}(\text{ReLU}_1) \end{aligned} \tag{8}$$

We obtained a higher-level abstract representation by introducing a projection layer, denoted as ProjectedT_v. During this process, our objective was to reduce dimensionality and extract key sentiment features from the raw data.

Similarly, through the text audio branch, we acquired text features that are augmented by audio (AugmentedT_a) and projected features ProjectedT_a.

3.3.2. Contrastive Learning Task: IPCL and SPCL

In the cross-modal text augmentation module, we introduced two contrastive learning tasks: instance prediction-based contrastive learning (IPCL) and sentiment polarity-based contrastive learning (SPCL). These tasks are employed to overcome the assumption of feature space consistency and effectively utilize feature space information containing inter-sample correlations and modality interactions to enhance the performance of multimodal sentiment analysis.

Firstly, in order to learn more abstract and robust sentiment representations within the augmented text features, we drew inspiration from contrastive learning efforts [35,46] and proposed an instance-based contrastive learning task. Specifically, we employed the prediction of augmented features using projection features, a prediction process that is cross-branched. In other words, we predict AugmentedT_a using ProjectedT_v and AugmentedT_v using ProjectedT_a. We treat (ProjectedT_v, AugmentedT_a) and (ProjectedT_a, AugmentedT_v) as our (query/key) pairs. During training, the contrastive learning task adheres to the principle of instance discrimination, where the query and key from the same sample form positive pairs, and other instances within the batch are treated as negative samples. The contrastive learning loss is computed using the widely used InfoNCE loss [52]. The specific calculation process is where we first compute the contrastive learning loss for (ProjectedT_v, a = AugmentedT_a) and (ProjectedT_a, AugmentedT_v) as follows:

$$\mathcal{L}_{\text{IPCL 1}} = -\log \frac{\exp(\text{sim}(\text{ProjectedT}_v, \text{AugmentedT}_a)/\tau)}{\sum_{i=1}^{n} \exp\left(\text{sim}\left(\text{ProjectedT}_v, \text{AugmentedT}_a^{(i)}\right)/\tau\right)} \tag{9}$$

$$\mathcal{L}_{\text{IPCL 2}} = -\log \frac{\exp(\text{sim}(\text{ProjectedT}_a, \text{AugmentedT}_v)/\tau)}{\sum_{i=1}^{n} \exp\left(\text{sim}\left(\text{ProjectedT}_a, \text{AugmentedT}_v^{(i)}\right)/\tau\right)} \tag{10}$$

where n denotes the batch size, sim is a similarity calculation function, and τ is a temperature parameter used to control the scaling of similarity scores. Subsequently, the sum of the two losses above constitutes the loss $\mathcal{L}_{\text{IPCL}}$ for the instance prediction contrastive learning task.

$$\mathcal{L}_{\text{IPCL}} = \mathcal{L}_{\text{IPCL 1}} + \mathcal{L}_{\text{IPCL 2}} \tag{11}$$

By pulling the distances between the positive sample pairs that we selected closer and pushing the distances between negative sample pairs further apart, the instance prediction-based contrastive learning task primarily achieves the following functions. By taking the contrastive learning of (ProjectedT_v, AugmentedT_a) as an example, ProjectedT_v represents the high-dimensional information of visually augmented text features, and AugmentedT_a represents the auditory augmented text clues. Bringing the distances between positive samples closer implies the alignment of the spatial features of the text features containing different modal information, integrating information from different modalities during the process, and accomplishing implicit multimodal feature fusion, as well as retaining the common sentimental features of different modalities. Since we use projected features to predict augmented features, this process also promotes the augmented features to approach higher-dimensional abstract sentiment, enhancing the stability of sentimental expression in the augmented features. At the same time, contrastive learning also pushes the distances between the negative sample pairs further apart, making the sentimental expressions of different samples more discriminative in space.

Following this, we extended the contrastive learning paradigm to supervised tasks. In the sentiment polarity-based contrastive learning task (SPCL), we leveraged label information to guide the model in learning more discriminative sentiment features. Specifically, we categorized the data into three classes: positive, neutral, and negative, based on the sentiment polarity criterion. In selecting sample pairs, samples with the same sentiment polarity as the target sample are considered positive samples, whereas samples with different sentiment polarities are treated as negative samples. We conducted contrastive learning

on the projection features; it is noteworthy that within the same batch, ProjectedT$_v$ and ProjectedT$_a$ are treated as a set by following the above partitioning rules and participating in the contrastive learning loss computation. This process also employs InfoNCE to compute the loss, which is expressed by the following formula:

$$\mathcal{L}_{\text{SPCL}} = -\log \frac{\exp\left(\text{sim}\left(P_{\text{target}}, P_{\text{positive}}\right)/\tau\right)}{\sum_{i=1}^{n} \exp\left(\text{sim}\left(P_{\text{target}}, P_{\text{negative}}{}^{(i)}\right)/\tau\right)} \tag{12}$$

where $P_{\text{target}} \in \{\text{ProjectedT}_v, \text{ProjectedT}_a\}$ and P_{negative} represent a situation with a different sentiment polarity from P_{target} within the mentioned set. τ is a temperature parameter.

SPCL continuously utilizes label information to adapt the projected features to sentiment analysis domain tasks. During this process, we adjusted the feature space of the projection features for the two branches, which can be regarded as a form of clustering operation. Simultaneously, this operation effectively enhances the model's focus on challenging samples within the features.

In the SPCL task, the projected features learned to be more sentimentally sensitive expressions, and this enhancement was also reflected in the IPCL task by influencing the augmented text features through cross-projection. The two tasks complement and influence each other, working together in different ways to promote augmented text features to exhibit a spatial distribution of features that is more consistent with sentiment analysis.

3.4. Total Training Loss

Finally, the two text augmentation features, AugmentedT$_v$ and AugmentedT$_a$, were combined. The resulting sum, F_{out}, passes through a dropout and the fully connected layers to obtain our predicted sentiment score, $\hat{y}$.

$$F_{\text{out}} = \text{AugmentedT}_v + \text{AugmentedT}_a \tag{13}$$

$$\hat{y} = FC(\text{Dropout}(F_{\text{out}})) \tag{14}$$

The predicted sentiment score, $\hat{y}$, along with the true labels of the samples, y, were used to calculate the regression task loss. Specifically, we employed mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{15}$$

where n represents the batch size. By combining our two contrastive learning losses, the overall training loss for the model is given by the following:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{MSE}} + \beta\mathcal{L}_{\text{IPCL}} + \gamma\mathcal{L}_{\text{SPCL}} \tag{16}$$

Here, α, β, and γ are the weighted coefficients utilized to adjust the contributions of different loss terms to the total loss.

4. Experiments

In this section, we initially present the relevant information about the datasets, evaluation metrics, experimental details, and baseline models used to validate the effectiveness of TCMCL. Subsequently, we showcase specific experimental results, analyzing the performance of TCMCL on the datasets. We conducted ablation experiments and visualization studies, further scrutinizing the impact of each component on the overall model.

4.1. Datasets and Evaluation Indicators

We evaluated the proposed TCMCL framework on the widely used multimodal sentiment analysis datasets CMU-MOSI [53] and CMU-MOSEI [54]. These datasets encompass

rich monologues from YouTube videos containing expressions of sentiment in the text, visual, and audio modalities. CMU-MOSI comprises 2199 video segments, each associated with a sentiment score label in the range of −3 to +3, representing the sentiment from negative to positive. CMU-MOSEI is more extensive, including 23,453 video segments covering 1000 different speakers across 250 diverse topics, and it is annotated with sentiment scores and an additional six emotion category labels. The data from both datasets underwent manual screening, a facial feature extraction confidence check, and a forced alignment confidence check. This process ensured that each video clip included in the datasets contained high-quality tri-modal information.

We performed two tasks on the datasets: regression analysis and classification analysis. We calculated the mean squared error (MAE) and Pearson correlation coefficients (Corr) for the regression tasks between our predicted values and actual sentiment scores. Subsequently, we transformed the regression model output for the binary classification tasks, determining whether the sentiment tendency was positive or negative. We used binary classification accuracy (Acc2) and the F1 score (F1) as evaluation metrics. Furthermore, we extended sentiment analysis to a finer-grained seven-class classification task, utilizing seven-class accuracy (Acc7) to comprehensively evaluate the model's capability in handling complex sentiment classification. The left side of the separator "/" represents the criteria for negative and non-negative classification, and the right side represents the positive and negative classification criteria.

4.2. Experimental Details

All the experiments in this study were conducted in a consistent experimental environment to ensure the comparability and reliability of the results. The model was trained on two NVIDIA GeForce RTX 3090 GPUs, with the environment consisting of Python 3.7 + PyTorch 1.9.0. The TCMCL model utilized BERT-BASE. After feature extraction, the d_t, d_a, and d_v were 768, 74, and 47, respectively. In the final loss computation, the weighted coefficients α, β, and γ were set to 1, 0.05, and 0.1, respectively. Regarding the training settings, the model's learning rate was set to 5×10^{-5}, using the Adam optimizer with a cosine annealing learning rate schedule (with a warm-up). A total of 80 epochs were trained with a batch size of 32. Additionally, Section 3 mentions some of the other parameters in the model, the specific settings of which are shown in Table 1.

Table 1. Parameter settings.

Parameter	Value
Length of feature sequence: N	50
Number of heads of multi-attention mechanisms: i	5
Number of layers of attention units: K	1
Projection layer: Linear layer size	768 and 128
Temperature parameter: τ	0.7
Dropout rate	0.5

4.3. Baselines

In order to validate the performance of our proposed model in multimodal sentiment analysis tasks, we compared it with the current leading models:

TFN [14]: The tensor fusion network aggregates the interactions between uni-modal, bi-modal, and tri-modal data through tensor fusion, facilitating the end-to-end learning of dynamics within and between modalities.

MFN [16]: The memory fusion network introduces a novel neural architecture for multiview sequential learning, explicitly addressing both view-specific and cross-view interactions through a system of LSTMs, the delta-memory attention network (DMAN), and multiview gated memory.

MFM [55]: The multimodal factorization model decomposes the representation into multimodal-discriminative and modality-specific generative factors. This optimization aids in achieving joint generative-discriminative objectives and reconstructing missing modalities.

MulT [32]: The multimodal transformer employs directional pairwise cross-modal attention for cross-modal interactions, which attends to the interactions between multimodal sequences across distinct time steps and latently adapts streams from one modality to another.

MISA [23]: The modality-invariant and -specific representations project modalities into two distinct subspaces, offering a holistic perspective on multimodal data for sentiment prediction tasks.

MAG-BERT [30]: The multimodal adaptation gate for BERT introduces a multimodal adaptation gate to enable fine-tuned BERT and XLNet models to incorporate multimodal nonlinguistic data.

Self-MM [22]: The self-supervised multimodal model engages in joint training for both multimodal and uni-modal tasks by autonomously generating single-modal labels and adjusting weight strategies. It simultaneously acquires modality consistency and distinctiveness.

MMIM [21]: MultiModal InfoMax maximizes mutual information hierarchically within uni-modal input pairs and between multimodal fusion results and uni-modal inputs to preserve task-related information during multimodal fusion.

MIB [24]: The multimodal information bottleneck achieves efficient multimodal representation learning by leveraging the information bottleneck (IB) to filter out uni-modal noise, ensuring powerful and nonredundant representations.

MMLATCH [17]: Bottom-up, top-down fusion proposes a neural architecture that captures top-down cross-modal interactions by using a feedback mechanism in the forward pass during network training.

SPECTRA [56]: Speech–text dialog pre-training for understanding spoken dialog with ExpliCiT cRoss-Modal Alignment improves downstream task performance by designing novel temporal location prediction tasks and cross-modal response selection tasks.

The SPECTRA model is trained using text and audio only, whereas all the remaining baselines simultaneously use text, audio, and visual modes for multimodal sentiment analysis.

4.4. Experimental Results

Table 2 presents the experimental results of all the baseline models and our proposed TCMCL model based on the aforementioned datasets and MSA metrics. The results demonstrate that our model achieves state-of-the-art performance on the CMU-MOSI and CMU-MOSEI multimodal sentiment datasets, with improvements across all evaluation metrics. Compared with the best baseline, TCMCL demonstrates enhancements in various metrics on both the CMU-MOSI and CMU-MOSEI datasets, with the most significant improvements being observed in the binary classification task (right), showing increases of 0.9% and 0.2% in Acc2, with the F1 also increasing by 0.9% and 0.2%. In the seven-class classification tasks, TCMCL achieves improvements of 0.3% and 0.5% on the two datasets. The model also performs exceptionally well in regression tasks, with the MAE metrics decreasing by 0.017 and 0.017 and the Corr metrics increasing by 0.01 and 0.006, respectively.

Overall, the models leveraging the pre-trained language model BERT for textual processing (e.g., MISA, Self-MM, MMIM, and MIB) consistently outperform earlier approaches in terms of multimodal sentiment analysis (MSA) performance across these benchmarks. These experimental findings underscore the efficacy of BERT in enhancing text representation and affirm the superiority of a text-centric multimodal contrastive learning framework for sentiment analysis.

Table 2. Experimental results of TCMCL versus baseline models on the CMU-MOSI and CMU-MOSEI datasets. The symbol ↓ indicates better performance as the data decreases, whereas ↑ indicates improved performance as the data increases. The best results are highlighted in bold and the second best results are marked with underlines.

Model	CMU-MOSI					CMU-MOSEI				
	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑	Acc7 ↑	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑	Acc7 ↑
TFN	0.944	0.672	79.3/80.0	79.3/80.1	33.8	0.566	0.708	80.1/82.1	80.2/82.3	48.8
MFN	0.952	0.695	79.1/80.6	79.0/80.5	32.7	0.589	0.725	79.9/82.4	80.0/82.6	47.4
MFM	0.915	0.704	79.8/80.4	79.8/80.2	33.2	0.632	0.719	80.0/82.8	80.6/83.0	49.2
MulT	0.787	0.783	80.8/82.1	80.9/82.2	36.2	0.617	0.722	82.5/83.5	82.6/83.7	50.9
MISA	0.771	0.786	81.6/83.2	81.6/83.3	39.1	0.599	0.724	82.1/84.3	82.4/84.4	48.9
MAG-BERT	0.731	0.783	82.7/85.0	82.6/85.0	44.3	0.563	0.749	82.3/84.9	82.6/84.9	51.4
Self-MM	0.727	0.787	83.0/85.1	82.3/85.1	44.2	0.559	0.744	81.5/85.1	81.7/85.1	51.2
MMIM	0.729	0.782	83.0/85.3	83.0/85.2	44.4	0.556	0.753	82.0/85.3	82.4/85.2	51.6
MIB	0.723	0.769	82.8/85.3	82.8/85.2	42.6	0.584	0.741	82.0/84.4	81.9/84.3	51.9
MMLATCH	0.736	0.721	81.7/84.1	81.7/84.1	43.0	0.582	0.723	81.2/83.0	81.2/83.0	52.1
SPECTRA	0.721	0.790	83.1/85.8	83.1/85.8	44.7	0.551	0.749	82.2/85.6	82.1/85.5	52.3
TCMCL	**0.704**	**0.807**	**84.4/86.7**	**84.3/86.7**	**45.0**	**0.541**	**0.759**	**82.8/85.8**	**83.2/85.7**	**52.8**

4.5. Ablation Study

In order to investigate the roles and influences of different modules in TCMCL further, we conducted ablation experiments using the CMU-MOSI dataset as an example, focusing on several aspects outlined below.

4.5.1. Uni-Modal Versus Multimodal

One of the key principles of TCMCL is to prioritize text as the central modality and augment it by incorporating audio and visual modalities. In order to demonstrate the effectiveness of this text-centric approach, we reconstructed the model's performance on the MSA task under various uni-modal settings. We also reconstructed the model with audio and visuals as the central modalities. Notably, in uni-modal experiments, where no other modalities are involved, self-attention is utilized in the augmentation module, rendering the two contrastive learning tasks ineffective. Finally, we constructed a balanced model with three central strategies in parallel. Specifically, the model maps the outputs of the three central strategies through a linear layer to the same dimension as the text-centric output. These outputs are then summed to form the output of the entire balanced model, with the total loss of the model being the sum of the MSE loss and the three sets of contrastive learning losses. The specific experimental results of all the models are shown in Table 3.

The experimental results indicate that under the uni-modal setting, the text modality demonstrates significantly better performance. This aligns with the findings of numerous prior studies [23,32], reflecting the superiority of text over the other two modalities in MSA tasks and affirming the rationale behind the text-centric model design strategy. On the other hand, under the multimodal setting, all three modalities show notable performance improvements, particularly the audio-centric and visual-centric approaches, which exhibit reductions of 0.163 and 0.188 in the regression metric MAE, respectively, and enhancements of 5.6% and 4.9% in the binary classification metric Acc2 (right). This significant progress is also attributed to the incorporation of textual information, enhancing the representation of sentiment features. Among the three central strategies, the text-centric multimodal contrastive learning (TCMCL) approach undoubtedly performs the best in MSA tasks. Compared with the other two strategies, audio-centric and visual-centric, TCMCL shows respective increases of 3.7% and 2.3% in Acc2 (right).

Table 3. Experimental results of the model under uni-modal conditions with three central strategies. The symbol ↓ indicates better performance as the data decreases, whereas ↑ indicates improved performance as the data increases. The best results are highlighted in bold.

	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑	Acc7 ↑
Text	0.793	0.769	81.9/83.8	81.7/83.8	41.0
Audio	0.971	0.667	76.5/77.4	76.5/77.5	30.2
Visual	0.944	0.723	78.0/79.5	78.0/79.6	32.8
Audio-centric	0.808	0.768	81.6/83.0	81.6/83.1	40.3
Visual-centric	0.756	0.780	82.5/84.4	82.6/84.4	42.4
Balance Model	0.747	0.778	82.2/84.9	82.2/84.8	41.6
Text-centric	**0.704**	**0.807**	**84.4/86.7**	**84.3/86.7**	**45.0**

Furthermore, compared with the balance model, the text-centric model continues to demonstrate superior performance. While the balance model incorporates contrastive learning tasks and cross-modal augmentation mechanisms, the inherent differences between the three centralized outputs inevitably introduce information interference and conflicts, hindering the accurate capture of sentiment features. In contrast, the text-centric model focuses on text information processing and enhances performance through contrastive learning and cross-modal augmentation, thus exhibiting better performance in sentiment analysis.

The performance variations observed in Table 3 among those models employing different modality design strategies underscore an important fact: the model design strategy profoundly influences the learning and recognition of sentiment information. Due to its representational advantage, the text modality occupies a paramount position in multimodal tasks, and the design of a text-centric approach leads to better MSA performance. In summary, the experiments demonstrate that a text-centric model design approach is the preferred solution for MSA tasks.

4.5.2. With or Without Contrastive Learning Tasks

In this study, we propose two contrastive learning tasks based on instance prediction and sentiment polarity to learn spatial representation information about augmented text features. In order to further investigate the contributions of these tasks to the model, we conducted a series of ablation experiments, as shown in Table 4.

Table 4. Results of ablation experiments on contrastive learning tasks in TCMCL. The symbol ↓ indicates better performance as the data decreases, whereas ↑ indicates improved performance as the data increases. The best results are highlighted in bold.

	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑	Acc7 ↑
w/o IPCL	0.731	0.785	83.0/85.5	83.2/85.5	43.4
w/o SPCL	0.719	0.797	83.9/86.1	83.8/86.1	43.9
w/o CL	0.745	0.784	82.8/84.7	82.8/84.7	42.3
TCMCL	**0.704**	**0.807**	**84.4/86.7**	**84.3/86.7**	**45.0**

Initially, we independently removed instance prediction-based contrastive learning (IPCL), resulting in significant discrepancies in all metrics compared with TCMCL, demonstrating the importance of IPCL. After removing IPCL, the MAE metric increased by 0.027, and correlation decreased by 0.022 in the regression task, whereas in the classification task, both Acc2 (right) and F1 (right) decreased by 1.2%, and the seven-class accuracy Acc7 decreased by 1.6%. The consistent performance degradation can be attributed to the crucial role of IPCL in implicit multimodal alignment and fusion. Ignoring this task results in more noise and misunderstanding introduced by directly manipulating the two augmented text features, given the vastly different impacts of the two auxiliary modalities on text. Subsequently, we eliminated sentiment polarity-based contrastive learning (SPCL), which

led to some performance drops across all metrics, albeit slightly less than with the removal of IPCL. The most noticeable changes were in Acc2 and F1, both decreasing by 0.7%, which directly related to sentiment polarity recognition performance. This indicates that under the supervision of sentiment polarity labels, SPCL indeed enables the model to learn more sentiment-sensitive feature representations. Finally, when both contrastive learning tasks are removed simultaneously, and the model is trained solely using regression task loss, we observe a further decline in model performance. These results indicate that regardless of which contrastive learning task is omitted, the model's performance is affected, affirming the indispensability of both contrastive learning tasks in the model.

The ablation study in this subsection validates some additional insights. The previous subsection mentioned that the pure-text model used self-attention for feature augmentation and ran without contrastive learning. This subsection demonstrates the performance of the model without contrastive learning when utilizing cross-modal text augmentation. We find that in the absence of contrastive learning tasks, cross-modal text augmentation leads to improved model performance, with a 0.9% increase in Acc2 (right) in the binary classification task. This demonstrates that the augmentation in the text-centric model relative to pure text is not solely attributable to contrastive learning tasks, validating the effectiveness of augmenting text features through auxiliary modalities.

4.5.3. Feature-Level Attention Versus Sequence-Level Attention

In the third section, we extensively discussed the cross-attention mechanism employed in this study, emphasizing feature-level attention instead of traditional sequence-level attention. We designed a comparative experiment to demonstrate the superiority of feature-level attention in augmenting text features. In this experiment, to simulate the application of sequence-level attention, we adjusted the model structure to refrain from transposing the features F_t (text), F_a (audio), and F_v (visual). Instead, we passed the audio and visual features through their respective linear layers, maintaining their dimensions, d_a and d_v, and keeping consistency with the text dimension, d_t. Table 5 presents the experimental results for both attention mechanisms in detail.

Table 5. Experimental results of sequence-level attention and feature-level attention. The symbol ↓ indicates better performance as the data decreases, whereas ↑ indicates improved performance as the data increases. The best results are highlighted in bold.

	MAE ↓	**Corr ↑**	**Acc2 ↑**	**F1 ↑**	**Acc7 ↑**
Sequence-level attention	0.739	0.788	82.9/84.8	82.9/84.9	41.6
Feature-level attention	**0.704**	**0.807**	**84.4/86.7**	**84.3/86.7**	**45.0**

The results indicate that feature-level cross-attention outperforms traditional sequence-level attention across all metrics. We speculate that the difference may stem from the fact that during sequence attention, we capture correlations between different modality sequences, but these correlations may vary significantly across different samples. In contrast, the finer granularity of feature-level attention—achieved by focusing on which features are more effective—can better capture cross-modal sentiment feature interactions while preserving the contextual information among text sequences. Integrating auxiliary modality information into text features on a feature-by-feature basis can augment the text more effectively, thereby boosting the model.

4.6. Parameter Experiments

We conducted sensitivity experiments on the key parameters of the TCMCL model to examine their impact on model performance.

4.6.1. Different Attention Unit Layers

In the design of the Siamese network, the text feature augmentation encoder employs a stack of K attention units. In order to explore the impact of the number of units, K, on the

model, we evaluated the binary classification accuracy metrics for different K settings. Our primary evaluation metrics are Acc2 and F1. The specific data are illustrated in Figure 3.

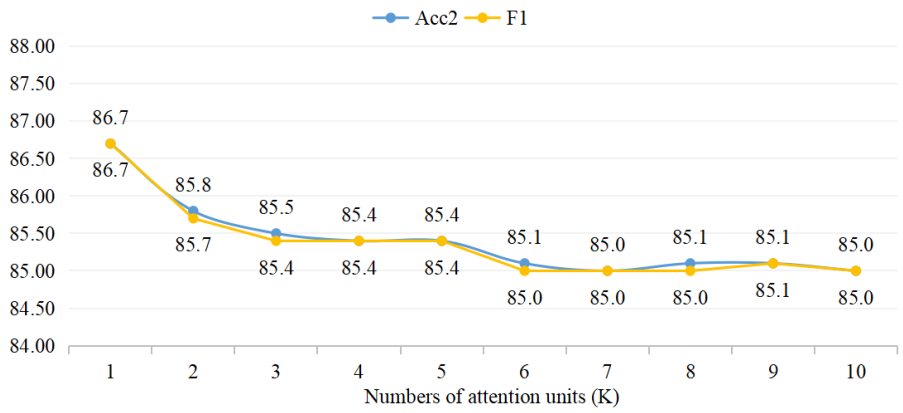


Figure 3. Experimental results of the TCMCL model with different numbers of attention units (K).

We comprehensively evaluated model performance across varying K values within the range of 1 to 10. The findings reveal a consistent decreasing trend in the metrics as K increases, eventually reaching a point of stabilization. Notably, the optimal performance is observed when K is set to 1.

4.6.2. Different Loss Weights

The experimental details subsection provides detailed descriptions of the weighting settings for model training loss. In order to thoroughly investigate the influence of the weighting coefficients α, β, and γ on model performance, we conducted a systematic experimental analysis of various weighting combinations. Specifically, the experiments fix α at 1 and vary β and γ to examine the relationship between α, β, and γ for quantification purposes. The evaluation metric for this experiment is binary classification accuracy (Acc2). Table 6 presents the specific values used in the experiment and the corresponding experimental results.

Table 6. Experimental results of TCMCL with different β and γ values. The best results are highlighted in bold.

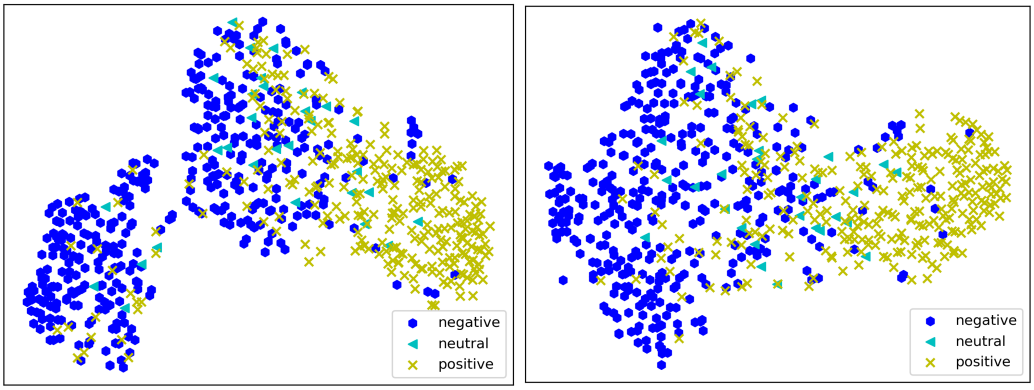
		β					
		0.05	0.1	0.15	0.2	0.25	0.3
γ	0.05	86.0	85.9	84.8	83.9	84.3	83.8
	0.1	**86.7**	85.4	84.6	84.0	84.2	83.6
	0.15	85.5	84.4	83.9	83.7	82.9	82.9
	0.2	84.7	84.1	84.0	83.5	82.7	82.6
	0.25	83.9	83.7	83.2	83.5	82.8	82.6
	0.3	84.0	83.2	83.0	82.4	82.5	82.3

The experimental results demonstrate that the model achieves optimal performance when β is set to 0.05 and γ to 0.1. It is evident from the table that the model performs better when β and γ are smaller. This phenomenon can be attributed to the primary objective of multimodal sentiment analysis, which is to minimize the discrepancy between the model's predicted scores and the true sentiment scores, and this is directly optimized by using the mean squared error (MSE) loss function. The incorporation of contrastive learning loss aims to further refine the distribution of the sentiment feature space for better multimodal information integration. However, assigning higher weights to the contrastive learning loss often shifts the model's focus toward adjusting spatial relationships between

samples, deviating from the primary task of multimodal sentiment analysis. This deviation leads to predicted scores drifting further from the ground truth, thereby impacting the model's performance.

4.7. Visualization Study

In order to validate the concept that the contrastive learning task can effectively utilize feature space information to learn superior sentiment features, we conducted a visual exploration using the test set of the CMU-MOSI dataset. Specifically, we performed T-SNE [57] dimensionality reduction visualization on the final feature outputs, F_{out}, of the model under conditions with and without contrastive learning tasks, as depicted in Figure 4a,b, respectively.



(**a**) Without contrastive learning (**b**) With contrastive learning

Figure 4. T-SNE visualization of the model feature output F_{out}.

From these figures, it can be observed that after the model performs contrastive learning tasks, the spatial distribution of the sentiment features exhibits a more pronounced clustering effect across different categories, showing more concentrated and structured clustering for the different sentiment categories. Particularly, within the negative sentiment, the clustering is no longer loose and disordered. The clustering centers of both the negative and positive sentiment features have also been pushed farther apart. Additionally, the distribution of neutral sentiment data in the space aligns better with the trend of sentiment variation. The visualization studies suggest that incorporating contrastive learning tasks can optimize the spatial distribution between samples and sentiment categories, facilitating a better understanding of the sentimental feature representation of the model.

5. Conclusions

This study introduces a novel text-centric multimodal contrastive learning framework for sentiment analysis (TCMCL). By diverging from previous studies, we emphasize the centrality of text in the framework's design, considering audio and visual features as supplementary augmentations to textual sentiment information. Two contrastive learning tasks were introduced to overcome the assumption of feature space consistency, directing attention to the information in the feature space. This approach not only captures implicit cross-modal interactions but also enhances the sentimental sensitivity of the features. Extensive experimentation on the CMU-MOSI and CMU-MOSEI datasets demonstrated the superiority of TCMCL, validating the effectiveness of the text-centric model design approach. Through visualization, we further elucidate the role of contrastive learning tasks in the model. In summary, this research provides new insights into the design strategy of such models. It is worth noting that there are currently various methods with which to augment the text modality using auxiliary modalities. In the future, we aim to explore multiple augmentation strategies to achieve outstanding performance in multimodal sentiment analysis.

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

Funding: This work was supported by the National Natural Science Foundation of China (NSFC; grant numbers: 62077027, 72091315 of 72091310), the Department of Science and Technology of Jilin Province, China (grant number: 20230201086GX), and the Industry University Research Innovation Fund of the Ministry of Education (grant number: 2022XF017).

Data Availability Statement: This study utilizes publicly available datasets from references [46,47].

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

References

1. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
2. Mullen, T.; Collier, N. Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 412–418.
3. Whitelaw, C.; Garg, N.; Argamon, S. Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, 31 October–5 November 2005; pp. 625–631.
4. Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 December 2003; pp. 427–434.
5. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [CrossRef]
6. Xia, R.; Liu, Y. Using denoising autoencoder for emotion recognition. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 2886–2889.
7. Deng, J.; Xia, R.; Zhang, Z.; Liu, Y.; Schuller, B. Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4818–4822.
8. Zhao, S.; Jia, G.; Yang, J.; Ding, G.; Keutzer, K. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Process. Mag.* **2021**, *38*, 59–73. [CrossRef]
9. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
10. Rosas, V.P.; Mihalcea, R.; Morency, L.P. Multimodal sentiment analysis of spanish online videos. *IEEE Intell. Syst.* **2013**, *28*, 38–45. [CrossRef]
11. Poria, S.; Cambria, E.; Hussain, A.; Huang, G.B. Towards an intelligent framework for multimodal affective data analysis. *Neural Netw.* **2015**, *63*, 104–116. [CrossRef] [PubMed]
12. Park, S.; Shim, H.S.; Chatterjee, M.; Sagae, K.; Morency, L.P. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2016**, *6*, 1–25. [CrossRef]
13. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.p.; Poria, S. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 6–15.
14. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1103–1114.
15. Barezi, E.J.; Fung, P. Modality-based Factorization for Multimodal Fusion. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Florence, Italy, 2 August 2019; pp. 260–269.
16. 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; Volume 32.
17. Paraskevopoulos, G.; Georgiou, E.; Potamianos, A. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual and Singapore, 23–27 May 2022; pp. 4573–4577.
18. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
19. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.

20. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
21. 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, 7–11 November 2021; pp. 9180–9192.
22. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 10790–10797.
23. 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 2020; pp. 1122–1131.
24. Mai, S.; Zeng, Y.; Hu, H. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Trans. Multimed.* **2022**, *25*, 4121–4134. [CrossRef]
25. Sun, L.; Lian, Z.; Liu, B.; Tao, J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2023**, *15*, 309–325. [CrossRef]
26. Chen, Q.; Huang, G.; Wang, Y. 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]
27. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
28. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
29. Yang, K.; Xu, H.; Gao, K. Cm-bert: Cross-modal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 521–528.
30. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the Conference, Association for Computational Linguistics, Meeting, Online, 5–10 July 2020; Volume 2020, p. 2359.
31. Kim, K.; Park, S. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. *Inf. Fusion* **2023**, *92*, 37–45. [CrossRef]
32. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Conference, Association for Computational Linguistics, Meeting, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.
33. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9729–9738.
34. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 1597–1607.
35. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15750–15758.
36. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef] [PubMed]
37. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; p. 1724.
38. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
39. Kumar, A.; Vepa, J. Gated mechanism for attention based multi modal sentiment analysis. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4477–4481.
40. Nguyen, D.; Nguyen, K.; Sridharan, S.; Dean, D.; Fookes, C. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Comput. Vis. Image Underst.* **2018**, *174*, 33–42. [CrossRef]
41. Zhang, Y.; Song, D.; Li, X.; Zhang, P.; Wang, P.; Rong, L.; Yu, G.; Wang, B. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Inf. Fusion* **2020**, *62*, 14–31. [CrossRef]
42. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Walker, M., Ji, H., Stent, A., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 2227–2237.
43. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. *Improving Language Understanding by Generative Pre-Training*; OpenAI: San Francisco, CA, USA, 2018.
44. Zeng, Y.; Li, Z.; Tang, Z.; Chen, Z.; Ma, H. Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Syst. Appl.* **2023**, *213*, 119240. [CrossRef]

45. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3733–3742.
46. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 21271–21284.
47. Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; Xu, W. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021.
48. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event, 7–11 November 2021; Association for Computational Linguistics (ACL): Kerrville, TX, USA, 2021; pp. 6894–6910.
49. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.
50. iMotions. Facial Expression Analysis. 2017. Available online: https://imotions.com/ (accessed on 20 October 2023).
51. 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 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 163–171.
52. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
53. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
54. 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 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
55. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning Factorized Multimodal Representations. In Proceedings of the International Conference on Representation Learning, New Orleans, LA, USA, 6–9 May 2019.
56. Yu, T.; Gao, H.; Lin, T.E.; Yang, M.; Wu, Y.; Ma, W.; Wang, C.; Huang, F.; Li, Y. Speech-Text Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 7900–7913.
57. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.