applied sciences

MDPI

Article

A Social Media Dataset and H-GNN-Based Contrastive Learning Scheme for Multimodal Sentiment Analysis

Jiao Peng [1], Yue He [1], Yongjuan Chang [1], Yanyan Lu [1], Pengfei Zhang [1], Zhonghong Ou [2,*] and Qingzhi Yu [3]

[1] State Grid Hebei Information and Telecommunication Branch, Shijiazhuang 050051, China; p2010015645@163.com (J.P.); 15133183154@139.com (Y.H.); chang02024@163.com (Y.C.); jackie_programmer@126.com (Y.L.); zhangpengfei9201@126.com (P.Z.)
[2] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China; yuqingzhi@bupt.edu.cn
* Correspondence: zhonghong.ou@bupt.edu.cn

Abstract: Multimodal sentiment analysis faces a number of challenges, including modality missing, modality heterogeneity gap, incomplete datasets, etc. Previous studies usually adopt schemes like meta-learning or multi-layer structures. Nevertheless, these methods lack interpretability for the interaction between modalities. In this paper, we constructed a new dataset, SM-MSD, for sentiment analysis in social media (SAS) that differs significantly from conventional corpora, comprising 10K instances of diverse data from Twitter, encompassing text, emoticons, emojis, and text embedded in images. This dataset aims to reflect authentic social scenarios and various emotional expressions, and provides a meaningful and challenging evaluation benchmark for multimodal sentiment analysis in specific contexts. Furthermore, we propose a multi-task framework based on heterogeneous graph neural networks (H-GNNs) and contrastive learning. For the first time, heterogeneous graph neural networks are applied to multimodal sentiment analysis tasks. In the case of additional labeling data, it guides the emotion prediction of the missing mode. We conduct extensive experiments on multiple datasets to verify the effectiveness of the proposed scheme. Experimental results demonstrate that our proposed scheme surpasses state-of-the-art methods by 1.7% and 0 in accuracy and 1.54% and 4.9% in F1-score on the MOSI and MOSEI datasets, respectively, and exhibits robustness to modality missing scenarios.

Keywords: multimodal sentiment analysis; datasets; contrastive learning; heterogeneous graph neural networks

Academic Editor: Stefan Fischer

Received: 14 November 2024
Revised: 4 January 2025
Accepted: 7 January 2025
Published: 10 January 2025

Citation: Peng, J.; He, Y.; Chang, Y.; Lu, Y.; Zhang, P.; Ou, Z.; Yu, Q. A Social Media Dataset and H-GNN-Based Contrastive Learning Scheme for Multimodal Sentiment Analysis. *Appl. Sci.* **2025**, *15*, 636. https://doi.org/10.3390/app15020636

1. Introduction

Multimodal sentiment analysis plays a crucial role in understanding emotional tendencies from diverse sources of information, such as text, images, and emojis. Its application in social scenarios and evaluation systems provides valuable insights into how users express emotions, which is vital for enhancing user experience and improving decision-making processes. For example, Chandrasekaran et al. [1] studied the application of multimodal sentiment analysis in social media and proposed specific methods for this domain. Zhang [2] and You [3] analyzed multimodal posts published on platforms like Weibo and YouTube to study user behavior. Kaur et al. [4] argued that stock prediction and product recommendation heavily rely on sentiment analysis to make better decisions.

The significance of multimodal sentiment analysis lies in its ability to integrate and analyze diverse emotional expressions across multiple modalities. In practical applications,

such as automatic processing of customer complaints, sentiment analysis can effectively summarize sudden and serious problems according to the emotions of users, and list the priority of the problems to be solved, so as to accelerate the processing efficiency. In recommendation systems, incorporating multimodal sentiment analysis can enhance the understanding of users' emotional preferences by analyzing sentiment in user reviews and interactions. This leads to more personalized recommendations. Furthermore, integrating sentiment analysis into intelligent customer service systems can improve the understanding of user needs, allowing for more accurate service responses by analyzing the emotional tone of user inquiries.

Some traditional multimodal sentiment analysis methods (will be introduced in related work) are based on examples of feature-level fusion and do not make full use of the complementary and correlated information between modalities. In order to improve the complementary and associated information between modalities, some recent works have introduced attention mechanisms or a graph neural network to establish associations between modalities. However, they still have two unanswered questions, missing modality and fusion strategy, which makes the exchange of information less effective.

Missing modality refers to the fact that not all modalities are included in the data, which affects the performance of multimodal tasks. Traditional completion methods use autoencoders or generative adversarial networks to recover missing modalities, but require a large amount of data and the quality is difficult to guarantee. Strategies such as contrastive learning or meta-learning can improve the generalization and robustness of completion methods, but still require analysis of the generated modalities, adding uncertainty.

Fusion strategy refers to how to effectively utilize the feature information between and within different modalities in multimodal data. Transformer-based methods use the attention mechanism to solve this problem, but do not fully consider the specificity and dynamics of the relationship between modalities. The graph neural network has also begun to be applied in the field of multimodal sentiment analysis, providing new ideas for multimodal alignment and fusion, and has achieved brilliant results. However, the graph neural network does not take into account the hierarchical relationship and weight distribution between modalities, resulting in insufficient and effective information exchange.

To address the issues mentioned above, we propose a graph neural network-based approach for multimodal sentiment analysis, which can adaptively learn hierarchical relationships and weight assignments among modalities, and can still maintain high performance when some modalities are missing. We use a contrastive learning strategy to achieve a weak alignment of data labels, splicing the completed labels onto node representations. We transplant the modality fusion problem into a heterogeneous graph, transform it into a node fusion problem, model the feature complementarity and dependence of multimodal data by automatically extracting meta-paths and graph convolutions, add virtual node aggregation modalities information, use H-GNNs to obtain node representations, and finally build a joint loss function for optimization, as shown in Figure 1.

We also built a new dataset for sentiment analysis in social media (SAS), which covers 10K pieces of multimodal data, including pictures, emojis, text, and other forms of information. Each piece of data is labeled with 16 emotional categories and 3 emotional polarities, using a variety of labeling methods to ensure the quality and consistency of the data. This dataset aims to address the diversity and complexity of data in social scenarios. Our main contributions and main problems solved are as follows:

- We have created a novel dedicated scene dataset, including 10K data, pictures, emojis, text, and other multimodal data, and carry out fine-grained annotations, including 16 kinds of emotional annotations and 3 kinds of polarity annotations, aiming to optimize a social scene sentiment analysis task. Compared with previous

multimodal emotion analysis datasets, SM-MSD incorporates a rich collection of emoji data, which enhances its comprehensiveness and informativeness. This makes SM-MSD particularly well suited for sentiment analysis tasks in social scenarios, providing higher-quality data support for research in this field. It is available at https://github.com/MR-YQZ/Social-Media-MultiSent-Dataset-SM-MSD- (accessed on 8 November 2024).

- We propose a multi-task framework that combines methods from heterogeneous graph neural networks (H-GNNs) and contrastive learning. The framework utilizes a contrastive learning strategy to achieve a weak alignment of data labels of different modalities, reconstruct semantic information of heterogeneous modalities, and complete emotional labels of missing modalities. At the same time, the framework adopts a multimodal data fusion and representation method based on heterogeneous graphs, uses meta-path extraction and graph convolutional networks to connect and aggregate information of different modalities, and introduces virtual nodes and attention mechanisms to fuse different information. The information on the meta-path can improve the fusion effect and representation ability.
- Our performance outperforms previous state-of-the-art models on multiple datasets, offering significant improvements and providing more accurate and reliable support for advancing research in this field.

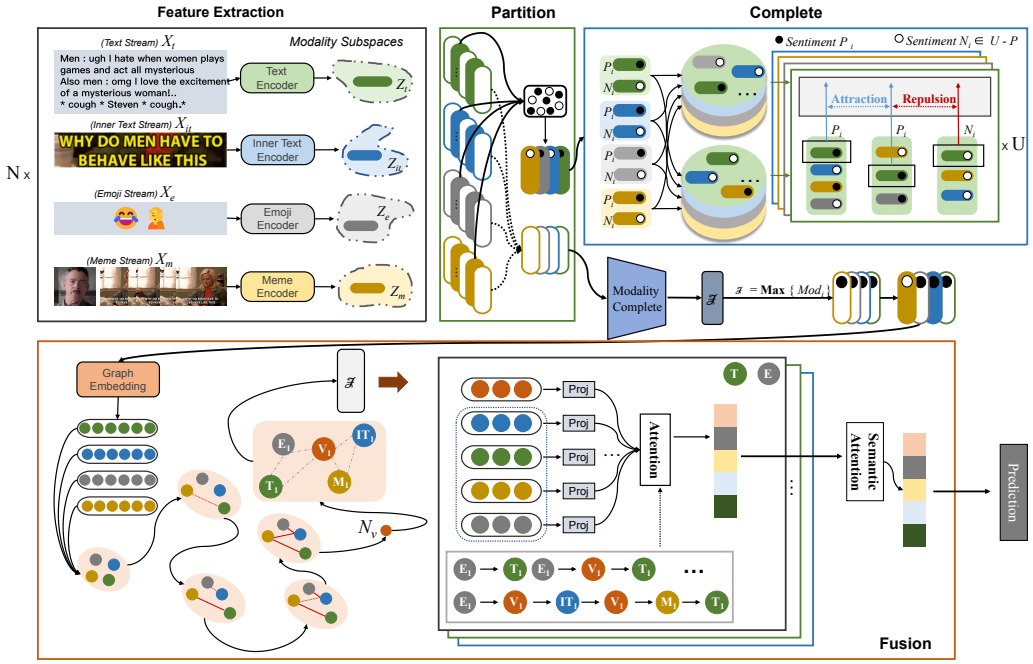

Figure 1. Our model first maps the data of different modes to different feature spaces, learns the representation of each emotion in each mode based on fine-grained modal labeling, generates the label of missing modes based on the existing modes, and finally aggregates the features of all modes through virtual nodes for each meta-path between modes as fusion features to complete the downstream classification task. The * in the picture of feature extraction is a symbol for a tone marker.

2. Related Works

2.1. Multimodal Sentiment Analysis Datasets

In order to adapt to various application scenarios, researchers have constructed datasets for multimodal sentiment analysis [5]. CMU-MOSI [6] is a multimodal sentiment analysis dataset. This dataset is a collection of YouTube monologues, providing sentiment annotations across modalities such as text, audio, and video. The sentiments

are marked as positive, negative, and neutral, and the value is between [−3, 3]. The CMU-MOSEI [7] dataset is an improvement on it. It is a large-scale emotion analysis and emotion recognition dataset. It mainly focuses on the speaker's facial expressions and contains three annotations from YouTube monologue videos, namely, emotion, emotion, and personality traits. Compared with CMU-MOSI, its sentiment annotations are more detailed. In social media platforms, in addition to text, audio, and video, emojis, memes, and other elements play an important role in expressing sentiment. However, these data are often overlooked in existing sentiment analysis datasets. Furthermore, the language used in social media is more colloquial, diverse, and influenced by internet and regional cultures, which existing datasets fail to fully reflect. For example, users frequently use abbreviations, slang, and tone markers on social media, and traditional datasets often lack the treatment of these features, leading to a reduced accuracy in sentiment analysis results. In addition, most of the existing datasets use coarse-grained annotation methods, which cannot meet the needs of fine-grained sentiment analysis.

2.2. Multimodal Sentiment Analysis Methods

Multimodal sentiment analysis mainly focuses on using multiple resources to predict human emotions [8]. The main challenge lies in modality fusion. Unlike traditional methods that only use a single modality, multimodal sentiment recognition aims to combine information from multiple sources [9] to improve the understanding and perception of human emotions. Previous multimodal approaches help to exploit complementary information across modalities. For example, Zadeh et al. [10] propose tensor fusion to explicitly capture interactions involving unimodal, bimodal, and trimodal data. Hazarika [11] proposes a novel framework, MISA, which projects each modality into two distinct subspaces, reducing inter-modal differences and capturing unique features.

In addition, existing methods can be broadly categorized based on their fusion techniques and modality representations:

- Tensor outer products or low-rank tensors to fuse features from different modalities: TFN and LMF [12];
- Factorization methods to decompose multimodal representations into different components: MFM [13] and FDMER [14];
- Attention mechanism to capture correlations and weights between different modalities: MISA;
- Canonical correlation analysis to coordinate different modalities into one hyperspace: ICCA [15].

Similarly, the node classification baselines commonly used in previous models are divided into the following categories according to the basic principles and application scenarios:

- Exploit the spectral properties of graphs to define graph convolution operations: GCN [16] and GAE [17];
- Exploit the spatial structure of the graph to define graph convolution operations: GAT [18], GIN [19], GTN [20], etc.;
- Handle heterogeneous graph data with many types of nodes and edges: HAN [21].

For node classification, HAN is the state-of-the-art (SOTA) model. HAN first extracts high-order semantic relations between different types of nodes in heterogeneous graphs according to the meta-path, and then uses the graph attention network to propagate and aggregate information on each meta-path, obtaining the node information expressed on each meta-path. Finally, the meta-path representation is fed into a hierarchical attention network, whose output is used for classification.

For multimodal sentiment analysis, the SOTA model is FDMER (Disentangled Representation Learning for Multimodal Emotion Recognition). FDMER is a feature

decomposition-based method for multimodal emotion recognition. FDMER first extracts public and private features from text, audio, and visual modalities, and then obtains effective multimodal representations. Finally, the multimodal representation is fed into a Canonical Correlation Analysis (CCA) network, whose output is used for prediction. The above methods are based on examples of feature-level fusion and do not make full use of the complementary and correlated information between modalities. In order to improve the complementary and associated information between modalities, some recent works have introduced attention mechanisms to establish associations between modalities. Tsai et al. [22] introduce a multimodal transformer to model unaligned language sequences, improving the integration of text and other modalities like images or audio. Garg et al. [23] discuss the use of multimodality in NLP applications and emphasize the need for effective fusion strategies across different data types. Zadeh et al. [24] further explore memory fusion for multi-view sequential learning, leveraging sequential data across different modalities for improved understanding. Hazarika et al. [25] present ICON, an interactive conversational memory network, which uses memory mechanisms to detect emotions across modalities in conversational contexts. However, these models can only describe the relationship between different data. The absence of a quantitative measurement will restrict the impact of subsequent interactions between modalities.

2.3. Incomplete Multimodal Data

In practical applications, multimodal data often have certain modes missing [26]. For example, on social media, users may only post text or pictures without speech. Multimodal task performance can be severely degraded by the missing number of modalities. For example, a piece of text might convey an angry tone, but if the speaker's facial expression is a playful smile, the context changes significantly. In this case, while the text may suggest anger, the facial expression and tone of voice may reveal a more nuanced sentiment, such as sarcasm or joking. Therefore, only through multimodal fusion can we effectively combine these modalities to make a more accurate prediction of sentiment, especially in social scenarios where emotions are often expressed in complex and contradictory ways.

In order to solve the problem of missing modes, some methods focus on the reconstruction of missing modes, that is, using existing modes to predict missing modes [27,28], but it is very complicated to use due to the huge amount of calculation of the generated model [29]. In a previous work, Ngiam proposed a variational autoencoder-based model that can reconstruct missing visual information from text and audio information, and uses adversarial learning to improve the reconstruction quality [30].

The above methods are examples based on fusion at the feature level and do not address the missing modality problem. In order to improve the problem of missing modes, some recent works try to use other existing modal information to reconstruct or predict the missing modal information [31], but these models require a large amount of labeled data, or can only deal with a single missing mode; this will impact the subsequent interactions between modalities.

2.4. Heterogeneous Graph Neural Networks

The heterogeneous graph neural network (H-GNN) is a graph neural network (GNN) method designed for heterogeneous graphs, and has been widely applied. For instance, Zeng et al. [32] propose an in-domain self-supervision method for heterogeneous graph convolution, enhancing multimodal sentiment analysis by exploiting the complementarity of different modalities. Similarly, Lu et al. [33] apply H-GNNs for aspect sentiment analysis, leveraging heterogeneous graphs to capture domain-specific information for better sentiment predictions. Other works, such as that by Linmei et al. [34], introduce heterogeneous

graph attention networks, which utilize attention mechanisms to weigh the importance of different modalities in semi-supervised settings, further improving performance in tasks like short text classification. These advancements are supported by foundational works in graph-based learning, such as GraphSAGE by Smith [35]. Shi et al. [36] also extend heterogeneous graph embeddings and attention mechanisms to improve sentiment analysis tasks, showcasing the effectiveness of heterogeneous graph-based models in handling multimodal data. In traditional homogeneous graphs, the nodes and edges must be of the same type, which prevents modeling different types of nodes from different modalities. However, in heterogeneous graphs, nodes and edges can come from different domains or have different semantics. H-GNN can simultaneously process information from multiple modalities (such as text, images, and audio) and efficiently fuse them through the graph structure. This allows H-GNN to effectively capture the complex relationships between different types of nodes [37–39], making it highly suitable for sentiment analysis tasks in multimodal scenarios.

In recent years, some researchers have extended graph neural networks [40–44] to the field of multimodal sentiment analysis. For example, Yang et al. propose a Multimodal Temporal Graph Attention Network (MTGAT) [45], which converts non-aligned multimodal sequences [46] into graphs with heterogeneous nodes and edges, using a multimodal temporal attention mechanism [47] and dynamic pruning and readout strategies to encode and decode graphs.

Although the above method is a typical example of HGNN-based multimodal sentiment analysis [10], it does not fully address the alignment, correspondence, refinement, and dynamization issues between modalities.

3. Dataset

3.1. Data Collection

It took about 3 months for the crawler to grab the tweets on the top page of Twitter. Selenium and a simulated sliding page, a fiddler capture tool, capture the request and response from the background of Twitter. Our data, SM-MSD, are inside the response. There is a json string in it, and there is a corresponding link to download the data. There are about 4–5 w every day, and 100 w of data are obtained.

We use the VGG19 pre-training model to roughly screen out non-emoji pictures, 95% of which are coarsely screened; manually fine-screen pictures and texts to obtain graphic-text dual-modal data; and manually remove 40%. Next, complex data processing is performed, using tools such as regular expressions to extract emoji from the text and using the PaddleOCR platform and manual correction to obtain the embedded text in the emoticon package. So far, all four modal data have been obtained. The process is illustrated in Figure 2. VGG19 is used when Sampling and Filtering "Noisy" refers to "Emoji", tools such as regular expressions are used when "Emoji" points to "Text", PaddleOCR is the part of Sampling and Filtering indicated by dotted lines, and the small man represents the artificial part.

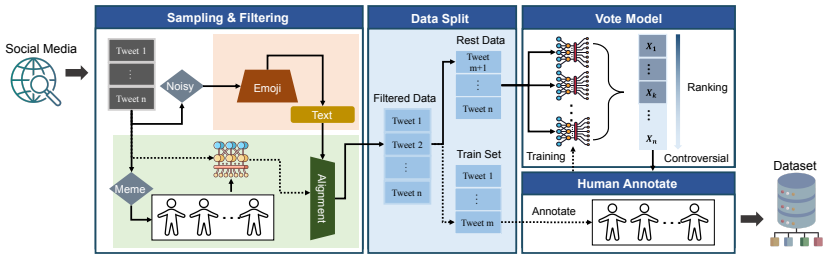

Figure 2. The process of collecting, filtering, and labeling our datasets.

3.2. Feature Extraction

Text and emoji. Since emoji can be represented by a special token, we regard emoji processing as text processing. In order to extract the features of text data, we use the "paraphrase-MiniLM-L6-v2" model provided by the SentenceTransformer library, that is, use and train Good BERT models to encode text.

Text in meme. We use PaddleOCR to extract text from pictures. It is an OCR tool library based on Paddle, which contains a variety of OCR scene application models, and supports Chinese and English number combination recognition, vertical text recognition, and other functions. Using PaddleOCR's OCR method, you can input the path of the image or the NumPy array, and output a list containing the text detection and recognition results. Each result contains two information of text content and confidence.

Meme. To classify memes by content, we divide them into static and dynamic categories based on whether they contain animation or not. For static memes, we use a 2D CNN to extract the visual features of the image. For dynamic memes, we use a motion vector-based method to select a few frames of images with large content changes, and then use the C3D model to learn the spatio-temporal features of the video.

3.3. Annotation

Build a multimodal fine-grained dataset, mark each modality separately, and mark the same data from multiple modal perspectives. Each piece of data has three types of annotations: text, picture, and comprehensive.

Labeling categories are divided into two categories according to polarity and emotion. The polarity is marked as positive, neutral, and negative. Choose 1 from 3, and the mood is marked as happy Joy, sad Sad, afraid of fear, angry, surprised, disgusted, and confused Confused's 16 multiple choices.

The annotations were completed by 10 researchers with excellent English. In the polarity annotation, –1 means negative, 0 means neutral, and 1 means positive. The average value given by the 10 annotators is taken as the final annotation. The emotional annotations given by the 10 annotators are discarded according to their distribution, and the reserved values are the final annotations.

We use a semi-automated approach to improve the size and quality of the dataset, provide labeled data with an active learning strategy, and prioritize the most controversial data, that is, more valuable data. The principle can be explained as follows.

Suppose there are N pieces of unlabeled data, and there are M base models, and each base model can give a category prediction for each piece of data. We define a voting function $v(u_i)$, which is used to calculate the number of votes of different categories for each piece of data in all base models. We define a difference function $d(u_i)$, which is used to calculate the difference between the largest number of votes and the second-largest number of votes for each piece of data in all base models. Our goal is to select K pieces of data with the largest difference for labeling, that is, to solve the following optimization problem:

$$\max_{S \subseteq \mathcal{U}, |S|=K} \sum_{u_i \in S} d(u_i) \tag{1}$$

This problem can be approximated by a greedy algorithm, that is, each time the one with the largest difference is selected from the unlabeled data and added to the label set until reaching K.

3.4. Statistics and Analysis

For the labeled data, we use the method of merging similar labels to improve the label imbalance problem, and use the cleanlab tool to clean up dirty data. Finally, 10K graphic

data were obtained. The modal distribution is shown in Tables 1 and 2. The sentiment distribution of different modalities is shown in Figure 3, and the polarity distribution is shown in Figure 4.

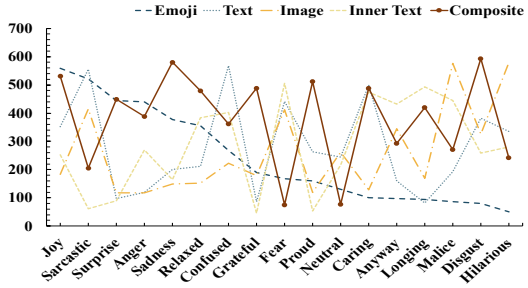

Figure 3. The sentiment distribution of different modalities.

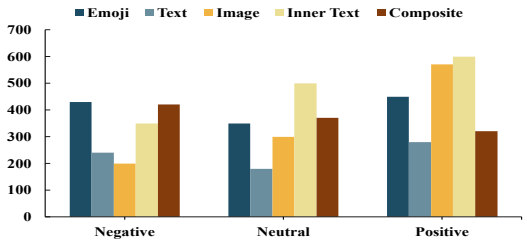

Figure 4. The polarity distribution of different modalities.

Table 1. Modality number.

Modality	Text	Emoji	Meme	Inner Text
Number	9.6 K	3.2 K	9.6 K	5.1 K
		Polarity Ratio		
Negative	34.29%	34.96%	18.69%	24.14%
Neutral	25.71%	28.46%	28.04%	34.48%
Positive	40.00%	36.59%	53.27%	41.38%
		Sentiment Ratio		
Joy	7.34%	13.54%	4.13%	5.24%
Sarcastic	11.58%	12.65%	9.32%	1.26%
Surprise	2.04%	10.78%	2.65%	1.84%
Anger	2.50%	10.68%	2.63%	5.59%
Sadness	4.19%	9.15%	3.35%	3.40%
Relaxed	4.42%	8.62%	3.41%	7.93%
Confused	11.85%	6.50%	5.01%	8.33%
Grateful	1.81%	4.61%	4.02%	0.95%
Fear	9.22%	4.08%	9.25%	10.46%
Proud	5.49%	3.88%	2.67%	1.12%
Neutral	5.11%	3.16%	5.88%	4.50%
Caring	10.41%	2.43%	2.87%	9.82%
Anyway	3.34%	2.38%	7.73%	8.95%
Longing	1.71%	2.31%	3.82%	10.21%
Malice	4.03%	2.09%	12.94%	9.20%
Disgust	7.97%	1.94%	7.32%	5.37%
Hilarious	6.99%	1.21%	13.00%	5.82%

Table 2. Another view of modality number.

Modality	Polarity	Sentiment
Overall	9.6 K	3.2 K
Text	34.29%	34.96%
Inner text	25.71%	28.46%
Emoji	40.00%	36.59%
Meme	40.00%	36.59%

3.5. Case Analysis

Figure 5 is a piece of data in SM-MSD, including Twitter text, emoji of laughing and crying, emoji of a man covering his face, meme in GIF form, and text inside a picture. Figure 5 shows the following from top to bottom: tweet, text in picture, two emojis, meme in GIF form.Tweets and emojis are saved in the data set as text, text in picture as picture, and meme in GIF form as GIF. From when annotating different modals, we can only see the corresponding modal data; that is, we cannot see the emoji information in the sentence when annotating text, and when annotating pictures, the text area is covered to ensure that the annotations of different modals are completely standalone; all modals are only visible when combined. In addition, the fine-grained annotations for each modality have both polarity and sentiment criteria.

Figure 5. Example of our dataset, SM-MSD. * is a symbol for a tone marker.

3.6. Data Source

This dataset was constructed by collecting publicly accessible data from Twitter. Data acquisition was performed using the Twitter API and includes publicly visible tweets and related information.

3.7. Data Processing

Data collection. Information in the dataset was obtained through automated scripts from Twitter. These data include tweet text, author information, and posting dates, among other details.

Data anonymization. Personal information of authors has been de-identified to protect their privacy. No specific user identification information is included.

Text cleaning. Text within the data has been processed and cleaned to remove sensitive or personal information while retaining information about topics and sentiments.

3.8. Data Usage Limitations

Legal purpose. This dataset is intended for legal purposes only, including research, analysis, and education. The use of data should comply with relevant legal regulations and ethical guidelines.

Privacy and ethics. Individuals using this dataset are expected to adhere to principles of user privacy and ethics. It should not be used for harassment, discrimination, privacy infringement, or any other unethical activities.

Data sharing. This dataset should not be used to create or perpetuate bias, inequality, racial, gender, or other inequality issues. Measures should be taken in research and analysis to mitigate these biases.

Transparency. Individuals or organizations using the dataset should provide transparent reports explaining how they are using the data and their research purposes.

Data protection. All individuals or organizations using the dataset should take measures to ensure data security to prevent unauthorized access or leakage.

4. Methodology

4.1. Model Overview

The goal of our model is to better fuse different modal representations of heterogeneous data and alleviate the performance degradation caused by missing modalities. The core idea is to use contrastive learning to generate pseudo-labels to "complete" the missing data of each modality, and then apply a two-level fusion method of node and meta-path to fuse the "completed" data. The overall structure is shown in Figure 1.

Feature extraction. The input of our model consists of four types of data: text, emoji, meme, and inner text. In the first part of the process, as shown in the upper left of Figure 1, feature extraction is performed on these four types of data. Through this step, the features of each modality are extracted and represented as embeddings.

Partition. The data are divided into two categories: complete modalities and missing modalities, as shown in the upper middle of Figure 1. This partition is based on whether data from a particular modality are missing or available. The incomplete modalities are separated and will undergo a different processing flow.

Complete. The complete modal data are used to train a model for modality completion, as shown in the upper-right section of Figure 1. The model learns to generate pseudo-labels and complete the missing parts of the data. This enables the system to predict the missing modality data, thus "complete" the missing data of each modality. When this step is completed, all the data have all the modes (original or "completed").

Prediction. A heterogeneous graph neural network (H-GNN) is used to model multimodal data and their relationships, as shown in the bottom part of Figure 1. This model aggregates features using meta-paths and adds a virtual node that associates all modalities, facilitating the fusion of information from different meta-paths. The fused node representations are then used to perform sentiment classification tasks.

The following details the core modules in the "complete" and "prediction" sections: "contrastive learning based on multimodal data completion" and "heterogeneous graph neural networks".

4.2. Contrastive Learning Based on Multimodal Data Completion

Sample division. For each data instance, we assess the presence or absence of modalities, categorizing the data into two sets: one consisting of complete modalities and the other of incomplete modalities. The data with complete modalities are utilized for training and testing the learning framework, while the trained model is employed to complete the missing modalities. Let the dataset with complete modalities contain N samples, each featuring three modalities: text, audio, and image. We denote the states of the ith sample as T_i, A_i, and I_i, respectively. Each sample is assigned a synthetic emotion label $Y_i \in 1, 2, \ldots, K$, where K represents the number of emotion categories [48]. Assuming that the image modal-

ity is absent [49], our objective is to complement the labels of the image modality using a contrastive learning approach [50].

Base learner. For each emotion $k \in 1, 2, \ldots, K$, we define a binary classifier $f_k(T, A)$, where the input consists of text and audio modalities, and the output provides a probability indicating whether the image modality corresponds to emotion k [51]. For each sample (T_i, A_i, Y_i), if $Y_i = k$, we set $Z_{ik} = 1$; otherwise, we set $Z_{ik} = 0$. This results in a binary label matrix $\mathbf{Z} \in 0, 1^{N \times K}$ for the image modality [52]. In this context, positive examples are defined as data that share the same modalities and synthetic sentiments as the current instance but differ in the missing modalities. For the case where the image modality is absent, we consider data whose images correspond to emotion A as positive examples for classifier A, while those whose images do not correspond to A are treated as negative examples.

Contrastive loss function. The goal of contrastive learning is to maximize the similarity between positive examples and minimize the similarity between negative examples. We utilize cosine similarity to evaluate this similarity. For each emotion k, we define the contrastive loss function L_k as follows:

$$L_k = -\frac{1}{N} \sum_{i=1}^{N} Z_{ik} \log \frac{\exp(\text{sim}(v_k(T_i, A_i), v_k(T_+, A_+)))}{\sum_{j=1}^{N} \exp(\text{sim}(v_k(T_i, A_i), v_k(T_j, A_j)))} \tag{2}$$

where (T_+, A_+) is a positive sample that shares the same modalities and synthetic emotions as (T_i, A_i) but differs in the image modality. The purpose of this loss function is to maximize the similarity between positive examples and minimize it between negative examples. To train the classifier $f_k(T, A)$, we minimize the contrastive loss function L_k using stochastic gradient descent. After training, the classifier $f_k(T, A)$ can be applied to predict whether any missing image modality data (T, A) correspond to emotion k. Specifically, we compute the output probability $p_k(T, A)$ of $f_k(T, A)$. This allows us to derive the emotional distribution $p_1(T, A), p_2(T, A), \ldots, p_K(T, A)$ for the image modality. The emotion associated with the maximum probability is taken as the predicted emotion for the missing data, represented as

$$\hat{Y}(T, A) = \text{argmax}_{k \in 1,2,\ldots,K} p_k(T, A) \tag{3}$$

The predicted label is concatenated as an additional dimension in the feature vector, thus ensuring that all modalities of each data instance are complete and enabling effective multimodal alignment.

4.3. Heterogeneous Graph Neural Networks

Heterogeneous graph neural networks (H-GNNs) are used for feature fusion of multimodal data, mapping data into heterogeneous graphs, combining artificial rules and automatic algorithms to obtain meta-paths, using graph convolutional networks to propagate and aggregate information; and increasing virtual nodes fuse information from different meta-paths, use an attention mechanism to calculate weights, use virtual nodes to represent the fusion of multimodal data, expand the number of meta-paths, and improve the fusion effect of long-distance related neighbors.

Graph embedding. This module describes how we convert each Twitter data into a heterogeneous graph node. Since different tweets contain different data types, each tweet contains a different number of nodes. A heterogeneous graph is constructed for all data $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each node $v \in V$ has a type $\phi_v \in \Gamma_v$, and each edge $e \in E$ has a type $\phi_e \in \Gamma_e$. Γ_v and Γ_e are a node type set and an edge type set, respectively. Each node v also has a feature vector $x_v \in R^d$, where d is the feature dimension.

Generate reformed graph. Our dataset implements fine-grained labeling, that is, labeling all modalities of each piece of data, and there is also a labeling for the entire piece of data. The virtual node is defined as v^*, and there is an edge connected to all modal nodes, and the type of this edge is ϕ_{e^*}. The state information is fused and passed to other nodes, and its eigenvector x_{v^*} can be initialized to the average value of all modal node eigenvectors or other methods.

We define a meta-path P as a sequence of adjacent edge types, such as $P = \phi_{e1} \rightarrow \phi_{e2} \rightarrow \cdots \rightarrow \phi_{ek}$, which can be used to represent a semantic relationship between two nodes.

We define a heterogeneous graph neural network H as a function, which can map the feature vector x_v of each node v to a low-dimensional vector $h_v \in R^k$, where k is the embedding dimension, and its goal is to enable h_v to capture the structural and non-structural information of v.

Aggregation. The graph neural network has a unique aggregation idea; that is, nodes are aggregated according to the meta-path [53]. The fusion of different types of data is achieved by sequentially aggregating adjacent nodes to itself. We assume that the heterogeneous graph neural network H consists of L layers, and each layer is a process of information aggregation and transformation. Level l can be expressed as

$$h_v^{(l)} = f^{(l)}\left(x_v, \left\{h_u^{(l-1)} : u \in N_v\right\}\right) \tag{4}$$

where $h_v^{(l)}$ is the embedding vector of node v in layer l, $f^{(l)}$ is the information aggregation and transformation function of layer l, and N_v is the set of neighbor nodes of node v, including the virtual node v^*. We can define different neighbor node sets N_v^P according to different meta-paths P; for example, $N_v^P = \left\{u : u \in N_v, \phi_{(u,v)} = P\right\}$ represents the set of neighbor nodes connected with node v by meta-path P.

In this way, we can obtain the embedding vector $h_v^{(l)}$ of each node in each layer, and the output $h_v^{(L)}$ of the last layer is the final embedding vector for v. We can use this embedding vector for tasks like multimodal sentiment analysis [52,54].

4.4. Objective Optimization

Suppose that there are N multimodal data samples $\{x_i\}_{i=1}^N$, each sample x_i contains M modes $\{x_i^{(m)}\}_{m=1}^M$, the target task is y_i, and the loss function L is designed to measure the error between the virtual node feature z_i output by H-GNNs and the target task y_i. We use the following cross-entropy loss function [55]:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log f(z_i) + (1 - y_i)\log(1 - f(z_i)) \tag{5}$$

Among them, $f(z_i)$ is a logistic regression function, which is used to map the virtual node feature z_i to the $(0, 1)$ interval, indicating the probability that x_i belongs to a certain category. The logistic regression function is defined as

$$f(z_i) = \frac{1}{1 + \exp(-z_i)} \tag{6}$$

Explanation: The first term, $y_i \log f(z_i)$, corresponds to cases where the sample x_i truly belongs to the target class ($y_i = 1$). The closer $f(z_i)$ is to 1, the smaller the loss. The second term, $(1 - y_i)\log(1 - f(z_i))$, corresponds to cases where the sample x_i does not belong to the target class ($y_i = 0$). The closer $f(z_i)$ is to 0, the smaller is the loss.

5. Experiments

5.1. Datasets

Since our task is essentially a node classification task, we choose not only multimodal sentiment datasets but also node classification datasets in graph neural networks.

CMU-MOSI. CMU-MOSI is a multimodal sentiment analysis dataset containing 93 video comments; each comment has a feature representation of voice, text, and facial expression and a label of emotional polarity (positive, negative, or neutral). The MOSI dataset is commonly used in multimodal emotion recognition and multimodal sentiment analysis.

CMU-MOSEI. CMU-MOSEI is a multimodal emotion and sentiment analysis dataset that provides data of multiple modalities such as video, audio, text, and facial expressions. The CMU-MOSEI dataset is a dataset of video reviews, where each review is labeled with an emotional polarity and six basic emotions. The CMU-MOSEI dataset is commonly used in tasks such as multimodal emotion recognition, multimodal sentiment analysis, and multimodal machine learning.

5.2. Experimental Setup

We conduct all experiments on an Ubuntu 18.04.2 LTS server with an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz, 256 G RAM, and 8 NVIDIA GeForce RTX 3090-24GB, sourced from the Procurement Office of Beijing University of Posts and Telecommunications, Beijing, China. We implement our model with Python 3.7.6 and PyTorch 1.7.0 with the framework of Deep Graph Library (DGL). The baselines are implemented from the original released codes or from the official implementation by DGL. Since the operations between the modules are additive, the additional computational cost is minimal, only introducing a slight constant overhead compared with other models. For a dataset setup, we evaluated our method on two datasets, MOSI and MOSEI. The data of the datasets are divided into the training set, the test machine, and the verification set according to the fixed 8:1:1 ratio. Finally, ACC, Precision, Recall, and F1 are used to evaluate the performance of our model and other models.

5.3. Results

The results on CMU-MOSEI and CMU-MOSI datasets are shown in Table 3. In these two datasets, our model achieves the best performance in all indicators and exceeds the baseline. In the results, we can see that our model is a structural attention fusion model, which performs better than the simple attention model. This is an encouraging result, because it means that the effect of model fusion is better after adding the structural information of data.

5.4. Ablation Studies

The importance of modality. We remove a single mode to explore the change of model performance. When the text mode is removed, the performance of the model drops significantly, which shows that the text mode is dominant in the task of multimodal sentiment analysis. When emoji and internal text are removed, a similar situation occurs, but the decline is smaller than that of text. A reasonable explanation is that image features are more complicated and redundant than text features. In addition, no matter which mode is removed, the performance of the model will decline to varying degrees, which shows that each mode has played its own role in multimodal sentiment analysis. The results are shown in Table 4.

Table 3. Comparison on multimodal sentiment analysis benchmark.

Model	MOSI			
	Acc_2 ↑	**Precision**	**Recall**	**F1 ↑**
Ours	85.12 ± 0.28	83.73 ± 0.47	82.61 ± 0.54	84.32 ± 0.29
TFN	80.04 ± 1.27	79.22 ± 1.46	78.63 ± 0.51	80.78 ± 0.99
LMF	82.53 ± 1.35	73.74 ± 1.39	75.59 ± 1.44	79.97 ± 0.67
MFM	81.73 ± 0.26	80.89 ± 0.75	79.42 ± 0.50	78.65 ± 0.08
ICCN	83.07 ± 1.29	82.25 ± 0.90	81.86 ± 0.71	83.04 ± 0.11
MISA	82.49 ± 0.34	83.57 ± 0.50	80.75 ± 0.78	81.69 ± 0.63
FDMER	83.68 ± 0.33	80.91 ± 0.52	73.19 ± 0.86	78.72 ± 0.27
Model	MOSEI			
	Acc_2 ↑	Precision	Recall	F1 ↑
Ours	86.14 ± 0.33	87.41 ± 0.45	88.29 ± 0.53	86.34 ± 0.28
TFN	82.52 ± 0.79	76.27 ± 1.32	80.48 ± 1.46	78.15 ± 0.33
LMF	82.06 ± 1.34	80.92 ± 0.40	80.02 ± 0.39	79.12 ± 0.30
MFM	84.45 ± 0.34	83.11 ± 0.78	82.28 ± 0.92	80.36 ± 0.13
ICCN	80.25 ± 0.08	83.01 ± 1.36	85.94 ± 0.63	82.22 ± 0.78
MISA	84.57 ± 0.32	80.95 ± 0.11	85.17 ± 0.76	82.31 ± 0.36
FDMER	86.17 ± 0.75	80.83 ± 0.99	83.01 ± 0.37	81.85 ± 0.09

Table 4. Results of studies on SM-MSD.

Model	Ours	Multimodal Sentiment Analysis			
	HCL	**TFN**	**LMF**	**ICCN**	**MISA**
Acc_3↑	73.24 ± 0.35	66.21 ± 0.57	69.36 ± 0.59	69.11 ± 0.64	70.04 ± 0.37
Precision	72.53 ± 0.65	65.86 ± 0.28	68.17 ± 0.09	67.73 ± 0.62	68.23 ± 0.45
Recall	70.31 ± 1.15	66.83 ± 0.74	67.62 ± 0.31	68.42 ± 0.48	69.61 ± 0.87
F1↑	74.01 ± 0.67	65.79 ± 0.97	67.84 ± 1.29	67.56 ± 0.47	66.47 ± 0.51

The importance of fine-grained modal annotation. We compare the effects of sentiment analysis using fine-grained modal labeling (labeling each mode and comprehensive mode) and coarse-grained modal labeling (labeling only comprehensive modes) to evaluate the effect of fine-grained modal labeling on improving the alignment and fusion between modes. We find that fine-grained modal labeling can bring better emotional analysis effect than coarse-grained modal labeling. This shows that fine-grained modal labeling can enhance the alignment and fusion between modes so that the model can make better use of multimodal information for emotion recognition. The results are shown in Table 4.

The importance of contrastive learning. We compare the efficiencies and complexities of contrastive learning and other methods. We find that contrastive learning can bring better emotional analysis effect than no modal completion.

Contrastive learning does not require generating additional modality data but directly uses existing data for learning. This saves time and computational resources for data generation and makes better use of available information.

Avoiding information loss: Through contrastive learning, you can minimize modality. Methods that generate modality data may introduce noise or information loss, while contrastive learning typically better preserves the features of the original data.

Computational efficiency: Generating modality data requires significant computational resources, whereas contrastive learning is usually more computationally efficient as it relies on existing data.

Practical applicability: Contrastive learning can be more easily applied in practical scenarios as it does not require generating additional data, which is an advantage for certain applications.

The results are shown in Table 5. Improved generalization: Contrastive learning helps the model learn relationships between multiple modalities, aiding better generalization to new data. Generating modality data methods often performs poorly when dealing with new data as they struggle to capture complex relationships between multimodal data.

Table 5. Results of ablation studies on SM-MSD.

Model	Acc_3 ↑	Precision	Recall	F1↑
Ours	**73.24 ± 0.32**	**72.53 ± 0.74**	**70.31 ± 0.58**	**74.01 ± 0.93**
Importance of modality				
w/o Text	53.41 ± 0.42	52.69 ± 0.78	49.27 ± 0.21	41.37 ± 0.45
w/o Emoji	69.21 ± 0.09	68.47 ± 0.37	67.94 ± 0.46	66.25 ± 0.84
w/o Meme	69.58 ± 0.37	67.98 ± 0.69	70.11 ± 0.83	70.29 ± 0.50
w/o Inner Text	70.52 ± 0.73	71.29 ± 0.05	68.43 ± 0.28	67.19 ± 0.16
Importance of fine-grained modal annotation				
w/o T-Label	71.47 ± 0.28	72.38 ± 0.73	69.83 ± 0.57	71.25 ± 0.49
w/o E-Label	71.76 ± 0.78	73.58 ± 0.27	71.04 ± 0.34	72.34 ± 0.18
w/o M-Label	72.09 ± 0.31	70.71 ± 0.03	71.59 ± 0.42	72.21 ± 0.25
w/o I-Label	71.94 ± 0.98	72.41 ± 0.85	69.73 ± 0.59	71.18 ± 0.51

The importance of modal completion. We randomly discard the modes to evaluate the effect of modal completion on improving data integrity and utilization. We find that modal completion can bring a better emotional analysis effect than no modal completion. This shows that modal completion can improve the integrity and utilization of data so that the model can better deal with the situation of missing modes. On both datasets, the effect of sentiment analysis with modal completion is significantly better than that without modal completion. The results are shown in Table 6.

Table 6. Results of contrastive learning on two benchmarks.

Model	MOSI		MOSEI	
	Acc_2 ↑	Precision	Recall	F1↑
HCL	85.12 ± 0.28	84.32 ± 0.29	86.14 ± 0.33	86.34 ± 0.28
20% data missing				
CL	85.08 ± 0.19	83.25 ± 0.43	85.72 ± 0.51	85.16 ± 0.37
w/o CL	80.15 ± 0.22	80.97 ± 0.49	83.91 ± 0.27	84.18 ± 0.31
40% data missing				
CL	84.65 ± 0.14	83.23 ± 0.26	81.37 ± 0.44	82.27 ± 0.53
w/o CL	75.58 ± 0.13	74.62 ± 0.54	75.14 ± 0.28	72.91 ± 0.37
60% data missing				
CL	73.89 ± 0.34	69.68 ± 0.55	70.28 ± 0.22	70.93 ± 0.41
w/o CL	59.97 ± 0.26	53.52 ± 0.31	56.37 ± 0.21	57.29 ± 0.48
80% data missing				
CL	62.18 ± 0.29	61.38 ± 0.21	66.31 ± 0.49	60.41 ± 0.39
w/o CL	48.29 ± 0.47	43.97 ± 0.22	50.26 ± 0.39	51.38 ± 0.31

6. Conclusions and Future Work

In this paper, a multimodal sentiment analysis method based on a graph neural network is proposed, which can adaptively learn the hierarchical relationship and weight distribution between modes and maintain high performance when some modes are missing. We transplant the modal fusion problem to heterogeneous graphs and transform it into a node fusion problem. We model the feature complementarity and dependence of multimodal data by automatically extracting meta-paths and graph convolution products, and add virtual nodes to aggregate modal information. On the MOSI and MOSEI datasets, accuracy improvements were 1.7% and 0%, respectively, and F1-score improvements were 1.54% and 4.9%, respectively, compared with the current best models. We also created a novel special scene dataset, including 10K data, images, and EMOJI text, and made fine-grained annotation, which is helpful in promoting the optimization of sentiment analysis tasks in social scenes.

Limitations: This work has certain limitations, including the exclusion of other types of multimodal data, such as video and audio. Additionally, we did not conduct a detailed analysis of the information conveyed by different meta-paths, nor did we explore more comparative learning strategies or loss function designs.

Future work: In the future, we plan to expand our dataset to include more diverse multimodal data sources. We will also explore more effective strategies for meta-path extraction and investigate novel graph convolution network structures to further enhance the performance of our method.

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

Funding: This research was funded by the Science and Technology Project of State Grid Hebei Electric Power Co., Ltd. (contract number SGHEXT00SJJS2310134).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: If you have any questions about this dataset or its usage, or require further information, please contact us. Our dataset is available at https://github.com/MR-YQZ/Social-Media-MultiSent-Dataset-SM-MSD- (accessed on 8 November 2024).

Conflicts of Interest: Authors Jiao Peng, Yue He, Yongjuan Chang, Yanyan Lu and Pengfei Zhang were employed by State Grid Hebei Information and Telecommunication Branch, Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chandrasekaran, G.; Nguyen, T.N.; Hemanth D, J. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1415. [CrossRef]
2. 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]
3. You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, Francisco, CA, USA, 22–25 February 2016; pp. 13–22.
4. Kaur, R.; Kautish, S. Multimodal sentiment analysis: A survey and comparison. In *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*; IGI Global: Hershey, PA, USA, 2022; pp. 1846–1870.

5. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
6. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
7. Bagher Zadeh, A.; 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; Gurevych, I., Miyao, Y., Eds.; pp. 2236–2246. [CrossRef]
8. 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, Montréal, QC, Canada, 18–22 October 2021; pp. 6–15. [CrossRef]
9. Jiang, M.; Ji, S. Cross-Modality Gated Attention Fusion for Multimodal Sentiment Analysis. *arXiv* **2022**, arXiv:2208.11893.
10. 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, 7–11 September 2017; Palmer, M., Hwa, R., Riedel, S., Eds.; pp. 1103–1114. [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, Virtual, 12–16 October 2020; pp. 1122–1131.
12. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
13. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv* **2018**, arXiv:1806.06176.
14. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled Representation Learning for Multimodal Emotion Recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 1642–1651.
15. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8992–8999.
16. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
17. Kipf, T.N.; Welling, M. Variational graph auto-encoders. *arXiv* **2016**, arXiv:1611.07308.
18. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
19. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? *arXiv* **2018**, arXiv:1810.00826.
20. Yun, S.; Jeong, M.; Kim, R.; Kang, J.; Kim, H.J. Graph transformer networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32* , 11983–11993.
21. Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; Yu, P.S. Heterogeneous graph attention network. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2022–2032.
22. 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 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Korhonen, A., Traum, D., Màrquez, L., Eds.; pp. 6558–6569. [CrossRef]
23. Garg, M.; Wazarkar, S.; Singh, M.; Bojar, O. Multimodality for NLP-Centered Applications: Resources, Advances and Frontiers. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Eds.; pp. 6837–6847.
24. 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 Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
25. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
26. Zhang, C.; Chu, X.; Ma, L.; Zhu, Y.; Wang, Y.; Wang, J.; Zhao, J. M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 14–18 August 2022; pp. 2418–2428. [CrossRef]
27. Rahate, A.; Walambe, R.; Ramanna, S.; Kotecha, K. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* **2022**, *81*, 203–239. [CrossRef]
28. Sun, W.; Ma, F.; Li, Y.; Huang, S.L.; Ni, S.; Zhang, L. Semi-supervised multimodal image translation for missing modality imputation. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 4320–4324.

29. 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]
30. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.
31. Tran, L.; Liu, X.; Zhou, J.; Jin, R. Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1405–1414.
32. 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]
33. Lu, G.; Li, J.; Wei, J. Aspect sentiment analysis with heterogeneous graph neural networks. *Inf. Process. Manag.* **2022**, *59*, 102953. [CrossRef]
34. Linmei, H.; Yang, T.; Shi, C.; Ji, H.; Li, X. Heterogeneous graph attention networks for semi-supervised short text classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4821–4830.
35. Hamilton, W.; Ying, Z,; Leskovec, J. GraphSAGE: Inductive Representation Learning on Large Graphs. In Proceedings of the 2017 Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1024–1034.
36. Shi, C.; Hu, B.; Zhao, W.X.; Philip, S.Y. Heterogeneous information network embedding for recommendation. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 357–370. [CrossRef]
37. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *Stat* **2017**, *1050*, 10-48550.
38. Zhang, C.; Song, D.; Huang, C.; Swami, A.; Chawla, N.V. Heterogeneous Graph Neural Network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 793–803.
39. Yu, L.; Shen, J.; Li, J.; Lerer, A. Scalable Graph Neural Networks for Heterogeneous Graphs. *arXiv***2020**, arXiv:2011.09679.
40. Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal learning with graphs. *Nat. Mach. Intell.* **2023**, **5**, 340–350. [CrossRef]
41. Gao, D.; Li, K.; Wang, R.; Shan, S.; Chen, X. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 12743–12753. [CrossRef]
42. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef] [PubMed]
43. Imran, M.; Yin, H.; Chen, T.; Huang, Z.; Zheng, K. DeHIN: A Decentralized Framework for Embedding Large-scale Heterogeneous Information Networks. *arXiv* **2022**, arXiv:cs.LG/2201.02757 . [CrossRef]
44. Bing, R.; Yuan, G.; Zhu, M.; Meng, F.; Ma, H.; Qiao, S. Heterogeneous graph neural networks analysis: A survey of techniques, evaluations and applications. *Artif. Intell. Rev.* **2023**, *56*, 8003–8042. [CrossRef]
45. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. Mtgat: Multimodal temporal graph attention networks for unaligned human multimodal language sequences. *arXiv* **2020**, arXiv:2010.11985.
46. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
47. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
48. Brinzea, R.; Khaertdinov, B.; Asteriadis, S. Contrastive Learning with Cross-Modal Knowledge Mining for Multimodal Human Activity Recognition. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8. [CrossRef]
49. Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; Faieta, B. Multimodal Contrastive Training for Visual Representation Learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 19–25 June 2021; pp. 6991–7000. [CrossRef]
50. Yang, M.; Yang, Y.; Xie, C.; Ni, M.; Liu, J.; Yang, H.; Mu, F.; Wang, J. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nat. Mach. Intell.* **2022**, **4**, 696–709. [CrossRef]
51. Liu, W.; Qiu, J.L.; Zheng, W.L.; Lu, B.L. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv* **2019**, arXiv:1908.05349.
52. Padi, S.; Sadjadi, S.O.; Manocha, D.; Sriram, R.D. Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models. *arXiv* **2022**, arXiv:2202.08974.
53. Liao, W.; Zeng, B.; Liu, J.; Wei, P.; Cheng, X.; Zhang, W. Multi-level graph neural network for text sentiment analysis. *Comput. Electr. Eng.* **2021**, **92**, 107096. [CrossRef]

54. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal sentiment detection based on multi-channel graph neural networks. 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, 1–6 August 2021; pp. 328–339.
55. Heaton, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618. *Genet. Program. Evolvable Mach.* **2018**, *19*, 305–307. [CrossRef]