# Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks

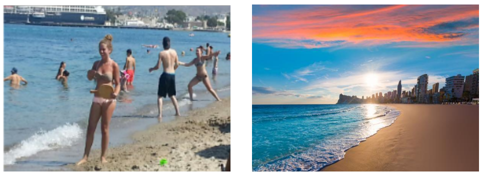
**Xiaocui Yang, Shi Feng, Yifei Zhang, Daling Wang**
School of Computer Science and Engineering, Northeastern University, China
yangxiaocui@stumail.neu.edu.cn,
{fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn

## Abstract

With the popularity of smartphones, we have witnessed the rapid proliferation of multimodal posts on various social media platforms. We observe that the multimodal sentiment expression has specific global characteristics, such as the interdependencies of objects or scenes within the image. However, most previous studies only considered the representation of a single image-text post and failed to capture the global co-occurrence characteristics of the dataset. In this paper, we propose Multi-channel Graph Neural Networks with Sentiment-awareness (MGNNS) for image-text sentiment detection. Specifically, we first encode different modalities to capture hidden representations. Then, we introduce multi-channel graph neural networks to learn multimodal representations based on the global characteristics of the dataset. Finally, we implement multimodal in-depth fusion with the multi-head attention mechanism to predict the sentiment of image-text pairs. Extensive experiments conducted on three publicly available datasets demonstrate the effectiveness of our approach for multimodal sentiment detection.

## 1 Introduction

The tasks of extracting and analyzing sentiments embedded in data have attracted substantial attention from both academic and industrial communities (Zhang et al., 2018; Yue et al., 2018). With the increased use of smartphones and the bloom of social media such as Twitter, Tumblr and Weibo, users can post multimodal tweets (e.g., text, image, and video) about diverse events and topics to convey their feelings and emotions. Therefore, multimodal sentiment analysis has become a popular research topic in recent years (Kaur and Kautish, 2019; Soleymani et al., 2017). As shown in Fig. 1, sentiment is no longer expressed by a pure modality in the multimodal scenario but rather by the combined expressions of multiple modalities (e.g., text, image, etc.). In contrast to unimodal data, multimodal data consist of more information and make the user's expression more vivid and interesting.



(a) We have a fun day on the beach! (*Positive*)

(b) We have a nice day on a deserted beach. (*Positive*)

Figure 1: Multimodal posts with global characteristics. Two posts express the user's positive sentiment from multimodal data that has global characteristics, including the *"have a fun/nice day"* phrase, the ocean scene, and the beach scene.

We focus on multimodal sentiment detection for image-text pairs in social media posts. The problem of image-text mismatch and flaws in social media data, such as informality, typos, and a lack of punctuation, pose a fundamental challenge for the effective representation of multimodal data for the sentiment detection task. To tackle this challenge, Xu et al. (2017; 2017) constructed different networks for multimodal sentiment analysis, such as a Hierarchical Semantic Attentional Network (HSAN) and a Multimodal Deep Semantic Network (MDSN). Xu et al. (2018) and Yang et al. (2020) proposed a Co-Memory network (Co-Mem) and a Multi-view Attentional Network (MVAN) models, respectively, introducing memory networks to realize the interaction between modalities.

The above methods treat each image-text post in the dataset as a single instance, and feature dependencies across instances are neglected or modeled implicitly. In fact, social media posts have specific global co-occurring characteristics, i.e., co-

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 328–339
August 1–6, 2021. ©2021 Association for Computational Linguistics

occurring words, objects, or scenes, which tend to share similar sentiment orientations and emotions. For example, the co-occurrences of the words "have a fun/nice day" and of the bright scenes "ocean/beach" in the two images in Fig. 1 imply a strong relationship between these features and positive sentiment. How to more effectively make use of the feature co-occurrences across instances and capture the global characteristics of the data remain a great challenge.

We propose a Multi-channel Graph Neural Networks model with Sentiment-awareness (MGNNS) for multimodal sentiment analysis that consists of three stages.

(i) **Feature extraction**. For text modality, we encode the text and obtain a text memory bank; for image modality, we first extract objects and scenes and then capture the image' semantic features from a multiview perspective.

(ii) **Feature representation**. We employ a Graph Neural Network (GNN) for text modality based on the global shared matrices, i.e., one text graph based on word co-occurrence is built based on the whole dataset. Specifically, we first connect word nodes within an appropriate small window in the text. After that, we update the node representation by itself as well as neighbor nodes. For image modality, it is believed that different views of an image, such as the beach (Scene view) and person (Object view) in Fig. 1(a), can reflect a user's emotions (Xu and Mao, 2017). The existing literature usually models the relationship between the scenes and objects within an image, failing to capture the rich co-occurrence information from the perspective of the whole dataset. In contrast, we explicitly build two graphs for scenes and objects according to the co-occurrences in the datasets and propose Graph Convolutional Network (GCN) models over the two graphs to represent the images. In general, to tackle the isolated feature problem, we build multiple graphs for different modalities, with each GNN acting as a channel, and propose a Multi-channel Graph Neural Networks (Multi-GNN) module to capture the in-depth global characteristics of the data. This multi-channel based method can provide complementary representation from different sources (George and Marcel, 2021; George et al., 2019; Islam et al., 2019).

(iii) **Feature fusion**. Previous studies usually directly connect multimodal representations, without considering multimodal interactions (Wang et al., 2020a; Xu, 2017; Xu and Mao, 2017). In this stage, we realize the pairwise interaction of text and image modalities from different channels through the use of the Multimodal Multi-head Attention Interaction (MMAI) module and obtain the fusion representation.

Our main contributions are summarized as follows:

- We propose a novel MGNNS framework that models the global characteristics of the dataset to handle the multimodal sentiment detection task. To the best of our knowledge, we are the first to apply GNN to the image-text multimodal sentiment detection task.
- We construct the MMAI module from different channels to realize in-depth multimodal interaction.
- We conduct extensive experiments on three publicly available datasets, and the results show that our model outperforms the state-of-the-art methods.

## 2 Related Work

### 2.1 Multimodal Sentiment Analysis

For convenience, multimodal polarity analysis and emotion analysis are unified to form multimodal sentiment analysis. Traditional machine learning methods are adopted to address the multimodal sentiment analysis task (Pérez-Rosas et al., 2013; You et al., 2016). Recently, deep learning models have also achieved promising results for this task. For the video dataset, Wang et al. (2020b) proposed a novel method, TransModality, to fuse multimodal features with end-to-end translation models; Zhang et al. (2020) leveraged semi-supervised variational autoencoders to mine more information from unlabeled data; and Hazarika et al. (2020) constructed a novel framework, MISA, which projects each modality to two distinct subspaces: modality-invariant and modality-specific subspaces. There is a massive amount image-text data on social platforms, and thus, image-text multimodal sentiment analysis has attracted the attention of many researchers. Xu et al. constructed different networks for multimodal sentiment analysis—HSAN (2017), MDSN (2017) and Co-Mem (2018). Yang et al. (2020) built an image-text emotion dataset, named TumEmo, and further proposed MVAN for multimodal emotion analysis.

## 2.2 Graph Neural Network

The Graph Neural Network has achieved promising results for text classification, multi-label recognition, and multimodal tasks. For text classification, a novel neural network called Graph Neural Network (GNN), and its variants have been rapidly developed, and their performance is better than that of traditional methods, such as Text GCN (Yao et al., 2019), TensorGCN (Liu et al., 2020), and TextLevelGNN (Huang et al., 2019). The GCN is also introduced in the multi-label image recognition task to model the label dependencies (Chen et al., 2019).

Recently, Graph Convolutional Network has been applied in different multimodal tasks, such as Visual Dialog (Guo et al., 2020; Khademi, 2020), multimodal fake news detection (Wang et al., 2020a), and Visual Question Answering (VQA) (Hudson and Manning, 2019; Khademi, 2020). Jiang et al. (2020) applied a novel Knowledge-Bridge Graph Network (KBGN) in modeling the relations among the visual dialogue cross-modal information in fine granularity. Wang et al. (2020a) proposed a novel Knowledge-driven Multimodal Graph Convolutional Network (KMGCN) to model semantic representations for fake news detection. However, the KMGCN extracted visual words as visual information and did not make full use of the global information of the image. Khademi (2020) introduced a new neural network architecture, a Multimodal Neural Graph Memory Network (MN-GMN), for VQA, which model constructed a visual graph network based on the bounding-boxes, which produced overlapping parts that might provide redundant information.

For the image-text dataset, we found that certain words often appear in a text post simultaneously, and different objects or scenes within an image have specific co-occurrences that indicate certain sentiments. We explicitly model these global characteristics of the dataset through the use of a multi-channel GNN.

# 3 Proposed Model

Fig. 2 illustrates the overall architecture of our proposed MGNNS model for multimodal sentiment detection that consists of three modules: the encoding module, the Multi-GNN module, and the multimodal interaction module. We first encode text and image input into hidden representations. Then, we introduce GNN from different channels to learn multiple modal representations. In this paper, the channels are the Text-GNN (TG) module, the Image-GCN-Scene (IGS) module, and the Image-GCN-Object (IGO) module. Finally, we realize the in-depth interactions between different modalities by multimodal multi-head attention.

## 3.1 Problem Formalization

The goal of our model is to identify which sentiment is expressed by an image-text post. Given a set of multimodal posts from social media, $P = \{(T_1, V_1), ..., (T_N, V_N)\}$, where $T_i$ is the text modality and $V_i$ is the corresponding visual information, $N$ represents the number of posts. We need to learn the model $f : P \to L$ to classify each post $(T_i, V_i)$ into the predefined categories $L_i$. For polarity classification, $L_i \in \{Positive, Neutral, Negative\}$; for emotion classification, $L_i \in \{Angry,\ Bored,\ Calm,\ Fear,\ Happy,\ Love,\ Sad\}$.

## 3.2 Encoding

**For text modality**, we first encode words by GloVe (Pennington et al., 2014) to obtain the embedding vector and then obtain the text memory bank, $M^t$, by BiGRU (Cho et al., 2014):

$$M^t = f_{BiGRU}(Embedding(T)), M^t \in \mathbb{R}^{L^t \times 2d^t}, \tag{1}$$

where $T$ is a text sequence, $L^t$ is the maximum length of a padded text sequence, and $d^t$ is the dimension of hidden units in the BiGRU.

**For image modality**, we extract image features from both the object and scene views to capture sufficient information. We believe that there are interdependencies between different objects or scenes in an image. To explicitly model this co-occurrence, we first extract objects $O = \{o_1, ..., o_{l^o}\}$ by YOLOv3 (Farhadi and Redmon, 2018), and extract scenes $S = \{s_1, ..., s_{l^s}\}$ by VGG-Place (Zhou et al., 2017). Finally, we obtain the object and scene memory banks with the pretrained ResNet (He et al., 2016). Thus, if an input image $V$ has a $448 \times 448$ resolution and is split into $14 \times 14 = 196$ visual blocks of the same size, then each block is represented by a 2,048-dimensional vector.

$$M^x = f^x_{ResNet}(V), M^x \in \mathbb{R}^{L^x \times d^x}, \tag{2}$$

where $x \in \{Object, Scene\}$, $L^x = 196$, and $d^x = 2,048$.

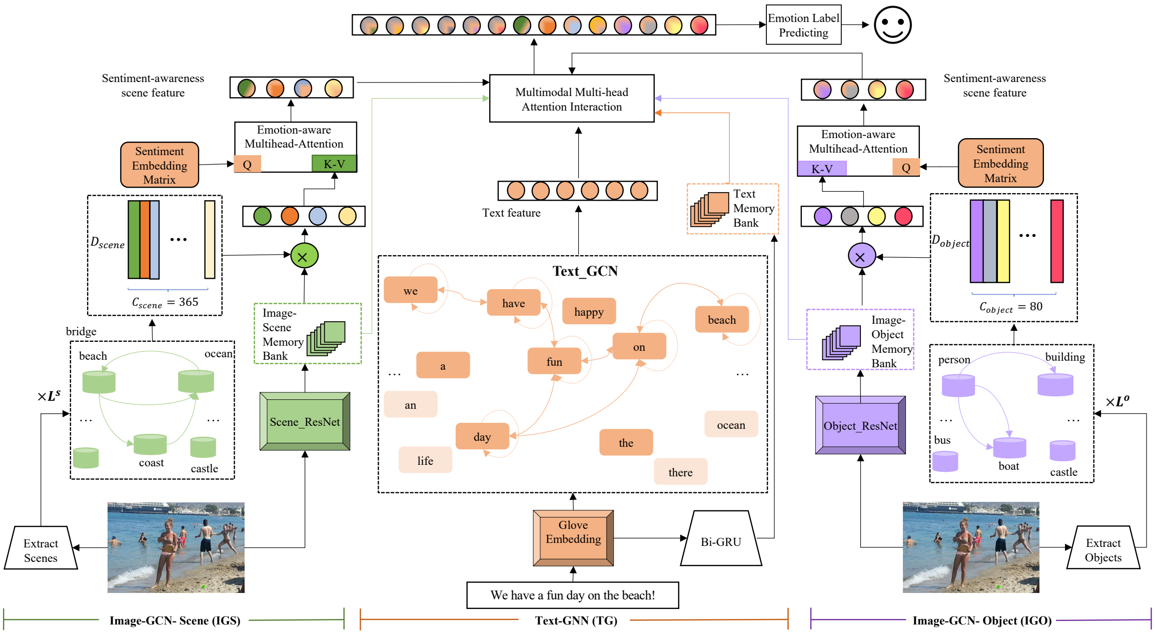


Figure 2: The framework of the proposed Multi-channel Graph Neural Networks with Sentiment-awareness (MGNNS) for multimodal sentiment detection. The channels are Text-GNN (TG) for text modality, Image-GCN-Scene (IGS) for image scene modality, and Image-GCN-Object (IGO) for image object modality. Note that we delete the stopwords during data preprocessing so that the words "a" and "the" do not have connections.

### 3.3 Multi-channel Graph Neural Networks

In this subsection, we present our proposed Multi-GNN module. As Fig. 2 shows, this module consists of the TG channel (middle), the IGO channel (right), and the IGS channel (left).

**Text GNN**: As shown in the middle of Fig. 2, motivated by (Huang et al., 2019), we learn text representation through the Text Level GNN. For text with $l^t$ words $T = \{w_1, ..., w_k, ..., w_{l^t}\}$, where the $k_{th}$ word, $w_k$, is initialized by glove embedding $r_k^t \in \mathbb{R}^d$, $d = 300$. We build the graph of the text-based vocabulary of the training dataset, which is defined as follows:

$$N^t = \{w_k | k \in [1, l^t]\}. \tag{3}$$

We build edges between $w_k$ and $w_j$ when the number of co-occurrences of two words is not less than 2.

$$E^t = \{e_{k,j}^t | w_k \in [w_1, w_{l^t}]; w_j \in [w_{k-ws}, w_{k+ws}]\}, \tag{4}$$

where $N^t$ and $E^t$ are the set of nodes and edges of the text graph, respectively. The word representations in $N^t$ and the edge weights in $E^t$ are taken from global shared matrices built based on vocabulary and the edge set of the dataset, respectively. That is, the representations of the same nodes and weights of the edges are shared globally. $e_{k,j}^t$ is initialized by point-wise mutual information (PMI) (Wang et al., 2020a) and is learned in the training process. $ws$ is the hyperparameter sliding window size, which indicates how many adjacent nodes are connected to each word in the text graph.

Then, we update the node representation based on its original representations and neighboring nodes by the message passing mechanism (MPM) (Gilmer et al., 2017), which is defined as follows:

$$A_k^t = \max_{j \in N_k^{ws}} e_{kj}^t r_k^t, \tag{5}$$

$$r_k^{t'} = \alpha r_k^t + (1 - \alpha) A_k^t, \tag{6}$$

where $A_k^t \in \mathbb{R}^d$ is the aggregated information from neighboring nodes from node $k - ws$ to $k + ws$, and $max$ is the reduction function. $\alpha$ is the trainable variable that indicates how much original information of the node should be kept, and $r_k^{t'} \in \mathbb{R}^d$ is the updated representation of node $k$.

Finally, we can calculate the new representation of text $T$ as follows:

$$T' = \sum_{k=1}^{l^t} r_k^{t'} \tag{7}$$

**Image GCN**: In this module, we explicitly model interdependence within $l^x$ scenes or objects by IGX, as shown on the left and right sides of Fig.

2, respectively. The graph of the image is defined as follows:

$$N^x = \{x_p | p \in [1, l^x]\}, \tag{8}$$

where $N^x \in \mathbb{R}^{C^x}$ is the set of nodes of IGX; $x\, or\, X \in \{Object, Scene\}$, $C^x = 80$ when $x = Object$, and $C^x = 365$ when $x = Scene$.

To build the edges of IGX, we first build the global shared co-occurrence matrix-based dataset:

$$E^x = \{e^x_{p,q} | p \in [1, l^x], q \in [1, l^x]\}, \tag{9}$$

where $E^x \in \mathbb{R}^{C^x \times C^x}$ is the co-occurrence matrix; edge weight $e^x_{p,q}$ indicates the co-occurrence times of $x_p$ and $x_q$ in the dataset.

Then, we calculate the conditional probability for node $p$ as follows:

$$P^x_{p,q} = e^x_{p,q} / N^x_p, when\ q \neq p \tag{10}$$

where $N^x_p$ denotes the occurrence times of $x_p$ in the dataset. Note that $P^x_{p,q} \neq P^x_{q,p}$.

As mentioned by (Chen et al., 2019), the simple correlation above may suffer several drawbacks. We further build the binary co-occurrence matrix:

$$B^x_{p,q} = \begin{cases} 1,\ if\ P^x_{p,q} \geq \beta \\ 0,\ if\ P^x_{p,q} \leq \beta \end{cases}, \tag{11}$$

where $\beta$ is the hyperparameter used to filter noisy edges.

It is obvious that the role of the central node is different from that of neighboring nodes, so we need to further calculate the weight of the edge:

$$R^x_{p,q} = \begin{cases} 1 - \gamma,\ if\ p = q \\ \gamma / \sum_{q=1}^{C^x} B^x_{p,q},\ if\ p \neq q \end{cases}, \tag{12}$$

where $R^x \in \mathbb{R}^{C^x \times C^x}$ is the weighted co-occurrence matrix, and hyperparameter $\gamma$ indicates the importance of neighboring nodes.

Finally, we input node $N^x$ and edge $R^x$ of the image into the graph convolutional network. Like in (Kipf and Welling, 2016), every layer can be calculated as follows:

$$H^x_{L+1} = h(\widehat{R^x} H^x_L W^x_L), \tag{13}$$

where $H^x_L \in \mathbb{R}^{C^x \times d^x}$, $H^x_{L+1} \in \mathbb{R}^{C^x \times d^{x'}}$, $W^x_L \in \mathbb{R}^{d^x \times d^{x'}}$, and $\widehat{R^x} \in \mathbb{R}^{C^x \times C^x}$ is the normalized representation of $R^x$; $h(\cdot)$ is a non-linear operation. When $L = 1$, $H^x_1$ is the word-embedding vector of $N^x$.

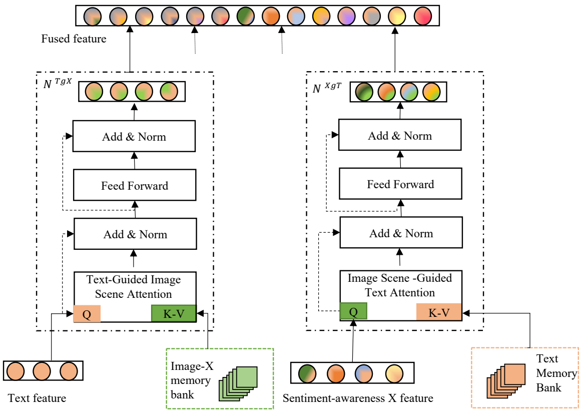


Figure 3: The MMAI module illustrates the process of multimodal interaction from four channels, $X \in \{Object, Scene\}$. We take the interaction process between text and image scene channels as an example to demonstrate this for convenience. The dotted arrows are the outputs of the other two channels after the interactions.

By stacking multiple GCN layers, we can explicitly learn and model the complex interdependence of the nodes. Then, we obtain the image representation with objects or scenes dependencies:

$$I^x = MaxPooling(M^x)(H^x_{L+1})^{\mathrm{T}},\ I^x \in \mathbb{R}^{C^x}. \tag{14}$$

But, we cannot capture the relationship between nodes and sentiments. Therefore, we learn the sentiment-awareness image representation through multi-head attention (Vaswani et al., 2017).

$$Att = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \tag{15}$$

$$\begin{aligned} EI^x &= MH(Q, K, V) \\ &= Concat(head_1, ..., head_H)W^O \\ where\ head_h &= Att(QW^Q_h, KW^K_h, VW^V_h), \end{aligned} \tag{16}$$

where $MH(\cdot)$ is multi-head attention; $W^Q_h \in \mathbb{R}^{d \times d_k}$, $W^K_h \in \mathbb{R}^{d_{model} \times d_k}$, $W^V_h \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{Hd_v \times d}$; and $H = 5, d_{model} = 300, d_k = d_v = 60$. $Q \in \mathbb{R}^{l^s \times d}$ is a sentiment embedding matrix built based on the label set $l^s = 3$ for polarity classification and $l^s = 7$ for emotion classification; $K = V = I^x W^I, W^I \in \mathbb{R}^{C^x \times d_{model}}, K, V \in \mathbb{R}^{d_{model}}$.

### 3.4 Multimodal Interaction

Motivated by the Transformer (Vaswani et al., 2017) prototype, we design a Multimodal Multi-head Attention Interaction (MMAI) module that can effectively learn the interaction between text modality and image modality by multiple channels, as shown in Fig. 3.

We employ the MMAI to obtain the Text guided Image-X representations and Image-X guided Text representations, $X \in \{Object, Scene\}$. For the **Text-guided Image-X attention**,

$$O_{N+1}^{TgX} = LN(MH(Q = H_N^{TgX}, K = V = M^x) + H_N^{TgX}), \tag{17}$$

$$H_{N+1}^{TgX} = LN(FFN(O_{N+1}^{TgX}) + O_{N+1}^{TgX}), \tag{18}$$

where $LN(\cdot)$ is layer normalization, and $FFN(\cdot)$ is the feed-forward network. When $N = 1$, $H_1^{TgX} = T'$, as in Eq. 7.

For the **Image-X-guided Text attention**,

$$O_{N+1}^{XgT} = LN(MH(Q = H_N^{XgT}, K = V = M^t) + H_N^{XgT}), \tag{19}$$

$$H_{N+1}^{XgT} = LN(FFN(O_{N+1}^{XgT}) + O_{N+1}^{XgT}), \tag{20}$$

when $N = 1$, $H_1^{XgT} = EI^x$, as in Eq. 16. For $MH$, $H = 4, d_{model} = 512, d_k = d_v = 128$. The fused multimodal representation is as follows: $R^m = [H_N^{TgO} \oplus H_N^{TgS} \oplus H_N^{OgT} \oplus H_N^{SgT}]$, where $\oplus$ is a concatenation operation.

### 3.5 Sentiment Detection

Finally, we feed the above fused representation, $R^m$, into the top fully connected layer and employ the softmax function for sentiment detection.

$$L^m = softmax(w^s R^m + b^s), L^m \in \mathbb{R}^{l^s}, \tag{21}$$

where $w^s$ and $b^s$ are the parameters of the fully connected layer.

## 4 Experiments

We conduct experiments on three multimodal sentiment datasets from social media platforms, MVSA-Single, MVSA-Multiple (Niu et al., 2016), and TumEmo (Yang et al., 2020), and compare our MGNNS model with a number of unimodal and multimodal approaches.

### 4.1 Datasets

**MVSA-Single and MVSA-Multiple** are two different scale image-text sentiment datasets crawled from Twitter[1]. **TumEmo** is a multimodal weak-supervision emotion dataset containing a large amount of image-text data crawled from Tumblr[2]. The statistics of these datasets are given in Appendix A; and for a fair comparison, we adopt the same data preprocessing method as that of Yang (Yang et al., 2020). The corresponding details are shown in Appendix B.

| Dataset | Train | Val | Test | All |
|---|---|---|---|---|
| MVSA-S | 3,608 | 451 | 452 | 4,511 |
| MVSA-M | 13,618 | 1,703 | 1,703 | 17,024 |
| TumEmo | 156,204 | 19,525 | 19,536 | 195,265 |

Table 1: Statistics of the different datasets.

### 4.2 Experimental Setup

| Parameter | MVSA-$*$ | TumEmo |
|---|---|---|
| Learning rate | $4e-5$ | $5e-5$ |
| $ws$ | 4 | 5 |
| Object-$\beta$ | 0.4 | 0.4 |
| Scene-$\beta$ | 0.3 | 0.5 |
| $\gamma$ | 0.2 | 0.2 |
| $L^x$ | 2 | 2 |
| $N^{TgX}$ | 1 | 1 |
| $N^{XgT}$ | 1 | 1 |

Table 2: Parameter settings of the different datasets.

We adopt the cross-entropy loss function and Adam optimizer. In the process of extracting objects and scenes, we reserve the objects with the probability greater than 0.5 and the top-5 scenes, respectively. The other parameters are listed in Table 2, $* \in \{Single, Multiple\}$. We use Accuracy (**Acc**) and F1-score (**F1**) as evaluation metrics. All models are implemented with PyTorch.

### 4.3 Baselines

We compare our model with multimodal sentiment models with the same modalities and the unimodal baseline models.

**Unimodal Baselines**: For text modality, **CNN** (Kim, 2014) and **Bi-LSTM** (Zhou et al., 2016) are well-known models for text classification tasks, and **BiACNN** (Lai et al., 2015) incorporates the CNN and BiLSTM models with an attention mechanism for text sentiment analysis. **TGNN** (Huang et al., 2019) is a text-level graph neural network for text classification. For image modality, **OSDA** (Yang

[1] https://twitter.com

[2] http://tumblr.com

| Modality | Model | MVSA-Single | | MVSA-Multiple | | TumEmo | |
|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | Acc | F1 |
| **Text** | CNN | 0.6819 | 0.5590 | 0.6564 | 0.5766 | 0.6154 | 0.4774 |
| | BiLSTM | 0.7012 | 0.6506 | 0.6790 | 0.6790 | 0.6188 | 0.5126 |
| | BiACNN | 0.7036 | 0.6916 | 0.6847 | 0.6319 | 0.6212 | 0.5016 |
| | TGNN | 0.7034 | 0.6594 | 0.6967 | 0.6180 | 0.6379 | 0.6362 |
| **Image** | OSDA | 0.6675 | 0.6651 | 0.6662 | 0.6623 | 0.4770 | 0.3438 |
| | SGN | 0.6620 | 0.6248 | 0.6765 | 0.5864 | 0.4353 | 0.4232 |
| | OGN | 0.6659 | 0.6191 | 0.6743 | 0.6010 | 0.4564 | 0.4446 |
| | DuIG | 0.6822 | 0.6538 | 0.6819 | 0.6081 | 0.4636 | 0.4561 |
| **Image-Text** | HSAN | 0.6988 | 0.6690 | 0.6796 | 0.6776 | 0.6309 | 0.5398 |
| | MDSN | 0.6984 | 0.6963 | 0.6886 | 0.6811 | 0.6418 | 0.5692 |
| | Co-Mem | 0.7051 | 0.7001 | 0.6992 | 0.6983 | 0.6426 | 0.5909 |
| | MVAN‡ | 0.7298‡ | 0.7139‡ | 0.7183‡ | **0.7038**‡ | 0.6553‡ | 0.6543‡ |
| | **MGNNS** | **0.7377** | **0.7270** | **0.7249** | 0.6934 | **0.6672** | **0.6669** |

Table 3: Experiment results of Acc and F1 on three datasets. ‡ represents the reproductive operation.

et al., 2020) is an image sentiment analysis model based on multiple views. Note that the SGN, OGN, and DuIG are variants of our model and rely only on image modality. **SGN** and **OGN** are the image graph convolutional neural networks based on scenes and objects for image sentiment analysis, respectively. **DuIG** is the image graph convolutional neural network with dual views, e.g., Object and Scene.

**Muiltimodal Baselines**: **HSAN** (Xu, 2017) is a hierarchical semantic attentional network based on image captions for multimodal sentiment analysis. **MDSN** (Xu and Mao, 2017) is a deep semantic network with attention for multimodal sentiment analysis. **Co-Mem** (Xu et al., 2018) is a co-memory network for iteratively modeling the interactions between multiple modalities. **MVAN** (Yang et al., 2020) is a multi-view attentional network that utilizes a memory network for multimodal emotion analysis. This model achieves state-of-the-art performance on image-text multimodal sentiment classification tasks.

### 4.4 Experimental Results and Analysis

The experimental results of the baseline methods and our model are shown in Table 3, where MGNNS denotes that our model is based on multi-channel graph neural networks[3].

We can make the following observations. First, our model (MGNNS) is competitive with the other strong baseline models on the three datasets. Note that the data distribution of MVSA-$*$ is extremely unbalanced. Thus, we reproduce the MVAN model with ACC and Weighted-F1 metrics instead of the Micro-F1 metric used in the original paper, which is more realistic. Second, the multimodal sentiment analysis models perform better than most of the unimodal sentiment analysis models on all three datasets. Moreover, the segmental indictors are difficult to capture for images owing to the low information density, and the sentiment analysis on the image modality achieves the worst results. Finally, the TGNN unimodal model outperforms the HSAN multimodal model, indicating that the GNN has excellent performance in sentiment analysis.

### 4.5 Ablation Experiments

We conduct ablation experiments on the MGNNS model to demonstrate the effectiveness of different modules. Table 4 shows that the whole MGNNS model achieves the best performance among all models. To show the performance of the Multi-GNN module, we replace the Text-GNN with the CNN, as well as the Image-GCN with the pretrained ResNet. The removal of the MMAI module (w/o MMAI) and Multi-GNN module (w/o MGNN) adversely affect the model results, which indicates that these modules are useful for multimodal sentiment analysis. By replacing the MMAI module with the CoAtt (Lu et al., 2016) module

[3] The source codes are available for use at https://github.com/YangXiaocui1215/MGNNS.

| Datasets | Model | Acc | F1 |
|---|---|---|---|
| MVSA-Single | w/o MGNN | 0.7010 | 0.6847 |
| | w/o MMAI | 0.7108 | 0.6879 |
| | +CoAtt | 0.7255 | 0.6986 |
| | w/o Scene | 0.7304 | 0.6988 |
| | w/o Object | 0.7034 | 0.6900 |
| | **MGNNS** | **0.7377** | **0.7270** |
| MVSA-Multiple | w/o MGNN | 0.7019 | 0.6752 |
| | w/o MMAI | 0.7128 | 0.6792 |
| | +CoAtt | 0.7210 | 0.6849 |
| | w/o Scene | 0.7170 | 0.6797 |
| | w/o Object | 0.7110 | 0.6848 |
| | **MGNNS** | **0.7249** | **0.6934** |
| TumEmo | w/o MGNN | 0.6553 | 0.6547 |
| | w/o MMAI | 0.6370 | 0.6347 |
| | +CoAtt | 0.6624 | 0.6606 |
| | w/o Scene | 0.6618 | 0.6593 |
| | w/o Object | 0.6592 | 0.6584 |
| | **MGNNS** | **0.6672** | **0.6669** |

Table 4: Ablation experiment results.

(+CoAtt), the model performance is found to be slightly worse than that of the MGNNS module. This further illustrates the importance of multimodal interactions and the superiority of the MMAI module. When one of the object views (w/o Object) or scene views (w/o Scene) is removed, the performance of the model declines, which indicates that both views of the image are effective for multimodal sentiment analysis.

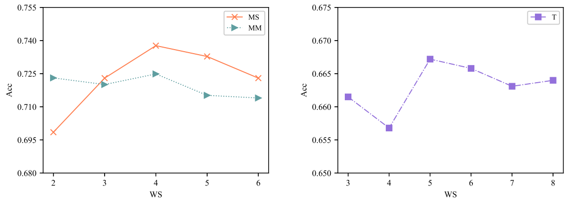
### 4.6 Transferability Experiment

In the Multi-GNN module, we build multiple graphs for different modalities based on the dataset. For different datasets, the graphs built by the unimodal model are different. However, can graph capture from one dataset (e.g., MVSA-Single) have positive effects on other datasets (e.g., TumEmo)? In this subsection, we will verify the transferability of the model through experiments.

As Table 5 shows, the following conclusions can be drawn: (i) Regardless of the modality, such as text or image, compared to introducing the graph constructed based on own dataset, the experimental results calculated based on graphs transferred from other datasets are worse. This is mainly because each dataset has unique global characteristics, the experimental results based on transferred graphs are slightly worse. (ii) However, due to the commonality of datasets when expressing the same emotions, the results of the transferred models are not completely worse. For example, the same scenes and objects can appear in different images in different datasets simultaneously for image modalities. Therefore, graphs from different datasets have transferability and can be used for other datasets. (iii) For different datasets, the experimental results of "X2Y-Text" are worse than those of "X2Y-Image". That is, the text graph has worse transferability. The reason for this may be that text graphs with various nodes are created based on the vocabulary of different datasets. Two situations in the transferred text graph will seriously affect the results: fewer nodes will lose information, and more nodes will provide redundant information. (iv) When the dataset gap is relatively wide, the transferability of text graphs is worse. For example, from the larger datasets transfer to the smallest dataset, including T2S-Text and M2S-Text, experimental results show a drop of 2.45% and 2.69%, respectively; from the smaller datasets transfer to the most largest dataset, including S2T-Text and M2T-Text, experimental results show a significant drop of 4.81% and 4.09%, respectively.

### 4.7 Hyperparameter Settings

**Hyperparameter** $ws$: To obtain adequate information from neighboring nodes in the TGNN, we conduct experiments under different settings for hyperparameter $ws$ in Eq. 4, the related results of which are shown in Fig. 4. The best $ws$ selection varies among different datasets since the average text length of TumEmo is longer compared to other data. The TGNN cannot obtain sufficient information from neighboring nodes with $ws$ values that are too small, while larger values may degrade the performance due to the redundant information provided by neighboring nodes.



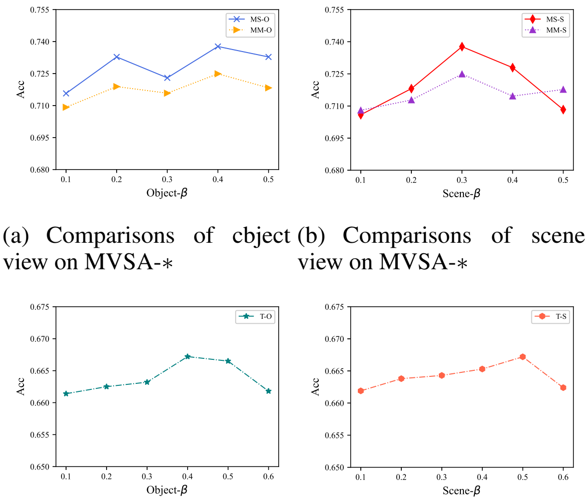
(a) Comparisons on MVSA-$*$ (b) Comparisons on TumEmo

Figure 4: Acc comparisons with different values of $ws$. MS is MVSA-Single, MM is MVSA-Multiple, and T is TumEmo.

| Model | MVSA-Single |  | Model | MVSA-Multiple |  | Model | TumEmo |  |
|---|---|---|---|---|---|---|---|---|
|  | Acc | F1 |  | Acc | F1 |  | Acc | F1 |
| M2S-Text | 0.7132 | 0.6985 | S2M-Text | 0.7146 | 0.6912 | S2T-Text | 0.6191 | 0.6202 |
| T2S-Text | 0.7108 | 0.6939 | T2M-Text | 0.7110 | 0.6752 | M2T-Text | 0.6263 | 0.6239 |
| M2S-Image | 0.7206 | 0.6901 | S2M-Image | 0.7177 | 0.6795 | S2T-Image | 0.6635 | 0.6611 |
| T2S-Image | 0.7255 | 0.7027 | T2M-Image | 0.7183 | 0.6848 | M2T-Image | 0.6625 | 0.6615 |
| **MGNNS** | **0.7377** | **0.7270** | **MGNNS** | **0.7249** | **0.6934** | **MGNNS** | **0.6672** | **0.6669** |

Table 5: Transferability experiment results of Acc and F1 on different datasets. S, M and T denote MVSA-Single, MVSA-Multiple, and TumEmo, respectively. For "Z" modality, "X2Y-Z" represents that the graph that is built based on the "X" dataset is transfered to the "Y" dataset, where Z ∈ {Text, Image}, X ∈ {MVSA-Single, MVSA-Multiple, TumEmo}, and Y ∈ {MVSA-Single, MVSA-Multiple, TumEmo}. For example, "M2S-Text" represents that the text graph that is built based on the MVSA-Multiple dataset is transferred to the MVSA-Single dataset.

**Hyperparameter** $\beta$: We vary the values of hyperparameter $\beta$ in Eq. 11 for the binary co-occurrence matrix from different views, the results of which are shown in Fig. 5. We find that the best $\beta$ value is different for different views in different datasets. For MVSA-$*$, the smaller $\beta$ value can reserve more edges to capture more information since the scene co-occurrence matrix is sparser than that in the object view. For TumEmo with a large amount of data, preserving the top-5 scenes produces many noise edges, so the value of scene-$\beta$ is greater than that of MVSA-$*$.



(a) Comparisons of cbject view on MVSA-$*$

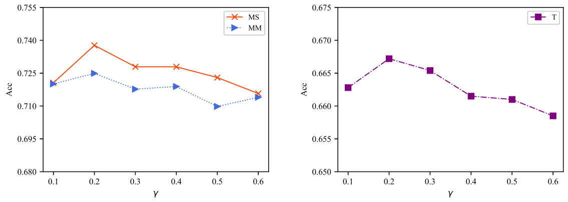
(b) Comparisons of scene view on MVSA-$*$

(c) Comparisons of object view on TumEmo

(d) Comparisons of scene view on TumEmo

Figure 5: Acc comparisons with different $\beta$ values.

**Hyperparameter** $\gamma$: As Fig. 6 shows, the model receives the best performance for the three datasets when $\gamma$ is 0.2. When $\gamma$ is smaller, the neighboring nodes do not receive enough attention; in contrast, their own information is not fully utilized.



(a) Comparisons on MVSA-$*$ (b) Comparisons on TumEmo

Figure 6: Acc comparisons with different $\gamma$ values.

## 5 Conclusions

This paper proposes a novel model, MGNNS, that is built based on the global characteristics of the dataset for multimodal sentiment detection tasks. As far as we know, this is the first application of graph neural networks in image-text multimodal sentiment analysis. The experimental results on publicly available datasets demonstrated that our proposed model is competitive with strong baseline models.

In future work, we plan to construct a model that adopts the advantages of the GNN and pre-trained models such as BERT, VisualBERT, and etc. We want to design a reasonable algorithm to characterize the quality of the objects and scenes selected from the image and further improve the representation ability of the model.

## Acknowledgments

The project is supported by the National Key R&D Program of China (2018YFB1004700) and by the National Natural Science Foundation of China (61772122, 61872074, U1811261).

## References

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*, pages 1724–1734.

Ali Farhadi and Joseph Redmon. 2018. Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition, cite as.*

Anjith George and Sebastien Marcel. 2021. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 16:361–375.

Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. 2019. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1263–1272.

Dan Guo, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, and Meng Wang. 2020. Iterative context-aware graph inference for visual dialog. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10055–10064.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for 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)*, pages 3442–3448.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Jumayel Islam, Robert E Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1355–1365.

Xiaoze Jiang, Siyi Du, Zengchang Qin, Yajing Sun, and Jing Yu. 2020. Kbgn: Knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1265–1273.

Ramandeep Kaur and Sandeep Kautish. 2019. Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 10(2):38–58.

Mahmoud Khademi. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7177–7188.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8409–8416.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, volume 29, pages 289–297.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference* *on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 973–982.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020a. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020b. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, pages 2514–2520.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402. ACM.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. 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*, pages 13–22. ACM.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2018. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, pages 1–47.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. *IEEE Access*, 8:22945–22954.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

# A Dataset

## A.1 MVSA-Single and MVSA-Multiple

The statistics for the MVSA-Simple and MVSA-Multiple datasets are listed in Table 1, showing that the various categories are highly unbalanced. MVSA-Single and MVSA-Multiple have different data distributions.

| **Dataset** | **Sentiment** | **Train** | **Val** | **Test** | **All** |
|---|---|---|---|---|---|
| MVSA-Simple | Positive | 2,146 | 268 | 269 | 2,683 |
| | Neutral | 376 | 47 | 47 | 470 |
| | Negative | 1,086 | 136 | 136 | 1,358 |
| | **All** | 3,608 | 451 | 452 | 4,511 |
| MVSA-Multiple | Positive | 9,054 | 1,132 | 1,132 | 11,318 |
| | Neutral | 3,526 | 441 | 441 | 4,408 |
| | Negative | 1,038 | 130 | 130 | 1,298 |
| | **All** | 13,618 | 1,703 | 1,703 | 17,024 |

Table 6: Number of Instances for Each Sentiment on the MVSA-$*$ Dataset.

| Emotion | Train | Val | Test | All |
|---|---|---|---|---|
| Angry | 11,635 | 1,454 | 1,455 | 14,544 |
| Bored | 25,826 | 3,228 | 3,229 | 32,283 |
| Calm | 14,487 | 1,811 | 1,811 | 18,109 |
| Fearful | 16,211 | 2,026 | 2,027 | 20,264 |
| Happy | 40,214 | 5,027 | 5,026 | 50,267 |
| Loving | 27,609 | 3,451 | 3,451 | 34,511 |
| Sad | 20,222 | 2,528 | 2,527 | 25,277 |
| **All** | 156,204 | 19,525 | 19,536 | 195,265 |

Table 7: Number of Instances of Each Emotion on the TumEmo Dataset.

### A.2 TumEmo

The statistics for the TumEmo dataset are listed in Table 2, containing a large number of image-text posts labeled by emotion.

## B Preprocessing Data

The text data contain many useless characters for sentiment analysis, such as URLs, stopwords, and punctuation. We need to preprocess text data to enhance the effectiveness of multimodal emotion detection. We perform data preprocessing as follows:

- remove the “URL”, as in“http://...”;
- remove the stopwords, such as “a, an, the, and etc. ”;
- remove the useless punctuation, including periods, commas, semicolons, etc;
- remove the hashtag and its content (#content); In particular, the TumEmo dataset uses #emotion as a weakly supervised label.
- remove the posts for which the text length is less than 3.