applied sciences

MDPI

Article

Global Local Fusion Neural Network for Multimodal Sentiment Analysis

Xiaoran Hu * and Masayuki Yamamura

Department of Computer Science, School of Computing, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama-shi 226-8502, Japan
* Correspondence: hu.x.ab@m.titech.ac.jp

Abstract: With the popularity of social networking services, people are increasingly inclined to share their opinions and feelings on social networks, leading to the rapid increase in multimodal posts on various platforms. Therefore, multimodal sentiment analysis has become a crucial research field for exploring users' emotions. The complex and complementary interactions between images and text greatly heighten the difficulty of sentiment analysis. Previous works conducted rough fusion operations and ignored the study for fine fusion features for the sentiment task, which did not obtain sufficient interactive information for sentiment analysis. This paper proposes a global local fusion neural network (GLFN), which comprehensively considers global and local fusion features, aggregating these features to analyze user sentiment. The model first extracts overall fusion features by attention modules as modality-based global features. Then, coarse-to-fine fusion learning is applied to build local fusion features effectively. Specifically, the cross-modal module is used for rough fusion, and fine-grained fusion is applied to capture the interaction information between objects and words. Finally, we integrate all features to achieve a more reliable prediction. Extensive experimental results, comparisons, and visualization of public datasets demonstrate the effectiveness of the proposed model for multimodal sentiment classification.

Keywords: multimodal sentiment analysis; fine-grained fusion; attention mechanism; deep learning

Citation: Hu, X.; Yamamura, M. Global Local Fusion Neural Network for Multimodal Sentiment Analysis. *Appl. Sci.* **2022**, *12*, 8453. https://doi.org/10.3390/app12178453

Academic Editor: Vincent A. Cicirello

Received: 20 July 2022
Accepted: 21 August 2022
Published: 24 August 2022

1. Introduction

The boom in mobile internet and smartphone access has made social networking an integral part of our daily lives. More and more people share their views and feelings through Twitter, Reddit, Weibo, and other social platforms, generating a large amount of social data. With the popularity of camera and video technology, data on these platforms have developed from a single text form to a combination of different media forms such as text, audio, and images. Massive multimodal data are formed in the form of a combination of text and images. Multimodal sentiment analysis aims to identify users' sentiment polarities, as well as their attitudes towards topics or events, from different forms of data. As the core field of social media analysis, sentiment analysis has received not only extensive attention from academia [1,2] but also has broad commercial application prospects, such as personalized advertising [3], opinion mining [4], and decision making [5], etc.

Compared with single-mode data, multimodal data contains more information and can easily express the users' real feelings. However, multimodal sentiment analysis is a challenging task. Different modal data have the possibility of containing different sentimental information, and the underlying features of these modalities have different dimensions and attributes. The traditional sentimental methods for multimodal sentiment analysis are based on handcrafted features [6] in each modality, and the performance highly relies on the quality of feature selection. Since the handcrafted features are usually designed with limited human knowledge, it is difficult to comprehensively describe the highly abstract features of sentiment to retain redundant emotional information and affect model

performance. With the development of machine learning and deep learning, researchers have proposed new sentiment analysis methods for unimodal and multimodal data, which have achieved promising results. Compared with handcraft features, convolutional neural networks [7,8], recurrent neural networks, and their variants [9] have been widely used for feature extraction, which can extract high-level features of each modality efficiently. In addition, benefitting from the introduction of large-scale datasets [10–12], many studies have begun using pre-trained models to analyze multimodal data better. Furthermore, deep-learning-based fusion methods have been proposed to explore the correlation among multimodal data [13–15].

The research about multimodal sentiment analysis is generally separated into two categories based on the dataset: video and image–text content. Some researchers leverage the effective methods [16–19] to understand the emotion of user-generated videos [20,21], which contain linguistic, visual, and acoustic information. Since the multimodal information parsed from the video is highly correlated, most approaches are centered around developing sophisticated fusion mechanisms to reduce the modality gaps between heterogeneous modalities. Other studies focus on the image–text data in social media. Visual and textual information have beneficial complementary properties in conveying information. This paper focuses on the text-image data from social media. Semantics in multimedia research remains challenging because of the complex relationship between text and images [22–25]. Following a general approach [24,25], the relationship between image and text is roughly divided into two parts: (1) Texts in multimedia mainly focus on the image content; the visual and textual information can be related based on the depicted or mentioned content. (2) Texts describe the feelings or evaluation of users; the image is semantically related to its corresponding text.

Intuitively, when the image content and the text have overlapping parts, people tend to focus on the objects or regions mentioned by the text in the image, thus forming visual and textual context information. These relevant contents often contain information that evokes specific human emotional responses, which is beneficial for multimodal data analysis. When there is no overlapping content between the image and the text, people will combine the general information of the image and the text to understand. The image and the text provide different information and complement each other to fully and effectively express the user's emotions. Some examples of posts which are collected from Twitter are shown in Figure 1. Since a boy, swimming pool, dog, etc., are mentioned in the text of Figure 1a, when we observe the picture, it is natural to pay attention to the local information of the image corresponding to the text, such as the boy, dog, etc., so as to combine the text with the corresponding image information to form contextual semantics. Such a correlation can be considered a fine-grained local cross-modal interaction. Different from (a), the images and texts in (b) do not directly express the overlapping information. The text expresses the user's feelings, experiences, and comments. The corresponding image supplements the information that the text does not represent, such as the scenery seen or the specific activities. When we see such text–image pairs, we tend to combine the image and text into a whole to form the semantics of the overall fusion. Taking into account these fusion relationships will benefit multimodal sentiment analysis.

According to the above analysis, this paper presents a global local fusion neural network, which considers the integrated fusion information as well as fine-grained fusion information between image and text for better understanding the visual and textual context and for effective sentiment analysis. Figure 2 shows the architecture of the proposed network, which consists of three sections: a global fusion module, local fusion learning, and an integration network. The model first obtains the modality-based overall fusion features from image and text, respectively, using an attention mechanism in the global fusion module. Then, in local fusion learning, the cross-modal module aligns image regions and sentence words, and fine-grained fusion builds the weight map for fusion features to enhance the interaction between the corresponding image and text content. Through these two modules, the coarse-to-fine interaction information is captured between

objects, scenes in images, and words in sentences adaptively. Finally, an integration network aggregates different levels of fusion information and comprehensively explores the multimodal sentiment.

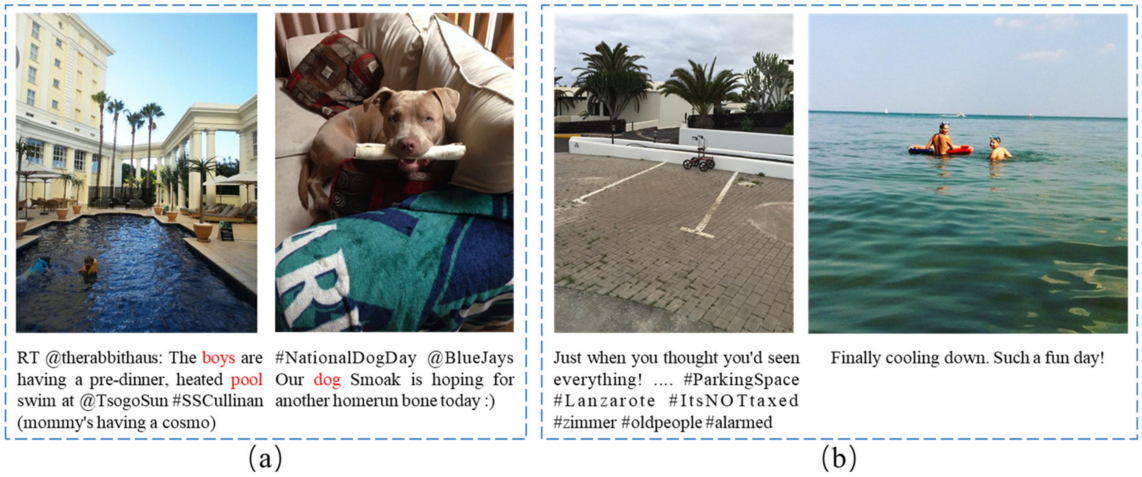


Figure 1. Examples of image–text posts from Twitter. (**a**) shows the fine-grained relation between image and text. (**b**) shows the global relation between image and text.

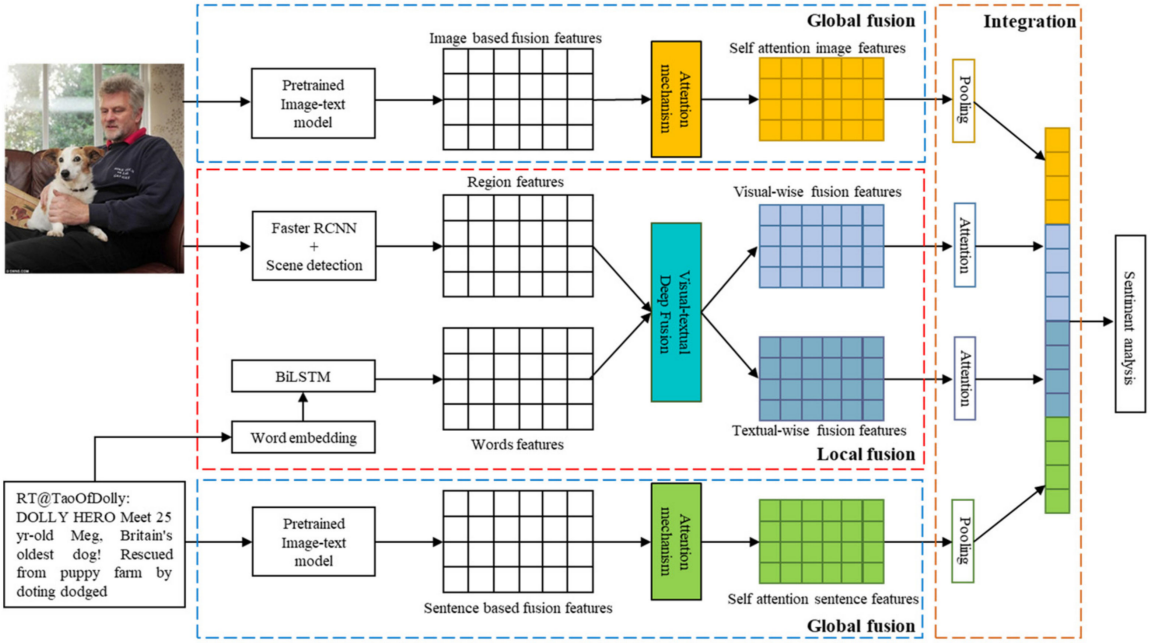


Figure 2. The overall framework of proposed model.

The main contributions of this paper are as follows:

1. A global local fusion neural network (GLFN) is proposed for the multimodal sentiment analysis task; the model captures fusion information at global and local levels to deal with different types of text–image relationships in social media and integrates all fusion information to achieve a more comprehensive prediction.
2. Symmetrical local fusion learning is introduced to effectively mine the modality-based corresponding information between image regions and text words. The fusion learning is coarse to fine; the cross-modal module aligns the features between image content and words as coarse fusion, and fine-grained fusion enhances the correlation between related content to strengthen the interaction.

Finally, extensive experiments are conducted in the public social media datasets, MVSA [26]. Comparisons with prior models are carried out to demonstrate the effectiveness of the proposed method in this paper.

The rest of this paper is organized as follows: Section 2 gives a summary of related works on sentiment analysis. Section 3 explains the detail of the proposed approach. Experiments, including baseline comparisons, ablation studies, and visualization, are reported in Section 4. Finally, the conclusion is summarized in Section 5.

2. Related Works

2.1. Unimodal Sentiment Analysis

The research for text sentiment analysis has been developed for many years. There are multiple levels of text sentiment analysis based on the length of datasets, including document-level, sentence-level, and aspect-level sentiment analysis. The methods used to analyze these textual data can be divided into two categories, the lexicon-based approach and the machine learning approach. The lexicon-based technique is extremely feasible for sentence-level analysis. Park and Kim [27] used a dictionary-based approach to build a thesaurus lexicon for sentiment tasks; the method adopted several dictionaries to collect a thesaurus and stored co-occurring words, which improved the classification accuracy without using human resources. Pang et al. [28] first applied machine learning methods, Naive Bayes, and maximum entropy for movie review sentiment analysis. Moreover, some methods used a combination of lexicon and machine learning. Borg and Boldt [29] first applied VADER with sentiment lexicon to provide the initial labels of emails and used LinearSVM to train the model and predict the sentiment of a not-yet-seen email, thereby preparing specific actions for customers who may have negative reactions. Recently, the deep neural network has been widely employed in natural language processing, including sentiment analysis. An attention-based bidirectional CNN-RNN model is proposed by [30], the model utilized two independent BiLSTM and GRU layers to capture past and future contexts in both directions, and an attention mechanism is applied to give weights to words for better performance.

Visual sentiment analysis primarily aims to explore the emotions associated with images. The first paper on image sentiment analysis was published in 2010 [31]. The authors built the correlation between the sentiment of image and the visual content and performed a discriminative feature analysis to predict the image sentiment. Borth et al. [32] proposed a large-scale visual sentiment ontology (VSO) based on psychological theories and web mining, which consists of many adjective–noun pairs (ANP), and then they trained the sentiment classifier with ANP outputs for the sentiment task. Yang et al. [33] extracted emotions from the images by leveraging all the related information, such as visual features, comments, and friendships; the model can distinguish the comments closely related to images' emotional expression. Deep learning approaches have also been used in image sentiment analysis. Chen et al. [34] introduced a convolutional neural network for the classification task. They performed fine-tuning training on the model and improved the results significantly. Song et al. [35] presented the network with an attention mechanism, which integrates visual attention into the CNN sentiment classification framework in an end-to-end manner. Wu et al. [36] employed a model which fused salient object information in the image with complete image information to predict the users' sentiment. The model shows that reasonably utilizing the local information could improve the model's performance.

2.2. Multimodal Sentiment Analysis and Pretrained Model

The multimodal sentiment analysis has attracted attention in recent years. Baecchi et al. [37] proposed a semi-supervised model, CBOW-DA-LR, which extended the CBOW model, learning textual and visual vector representations concurrently to build the sentiment polarity classifier. Hu and Flaxman [38] used fine-tuning Inception and word representation GloVe [39] combined LSTM to extract high-level visual and textual features and concatenated the features as input of dense layer for the sentiment task. Poria et al. [40]

utilized 1D CNN, RNN, and 2D CNN to obtain the features from the text, audio, and image and applied a multiple-kernel learning classifier to fuse the multimodal information for the sentiment classification. Many studies focus on the fusion method in the multimodal sentiment task. Zadeh et al. [41] proposed a tensor fusion network, which learned intra-modality and inter-modality dynamics end to end. Xu et al. [42] proposed a co-memory network with attention mechanism, which captured the interaction of image and sentence iteratively to conduct the sentiment analysis. Jiang et al. [43] proposed the FENet model, which included interactive information fusion to learn the fusion features and a specific information extraction to extract sentimental features for sentiment prediction. Hu et al. [44] introduced a two-stage attention-based fusion neural network to analyze textual–visual information for sentiment classification. Yang et al. [45] proposed a model based on the multi-view attention network, which combined object–text fusion and scene–text fusion to tackle the task. Li et al. [46] proposed a contrastive learning and multi-layer fusion network to detect sentiment. To build more effective correlations between image and text, object and scene extraction methods have been used to catch more details of the images. Zhu et al. [47] introduced an image–text interaction network to investigate the relationship between image and text for sentiment classification. Although the researchers obtained exciting results, the models did not consider the relationship between image and text and coarse-grained fusion.

Recently, with the sophisticated pre-training targets and huge parameters, large-scale pre-trained models have demonstrated significant performance in many fields. The early exploration of the natural language process (NLP) pre-trained model is a shallow network, such as Word2Vec [48] and GloVe [39], which can represent semantic meanings of words. With the development of NLP, Transformers [49] are proposed, and based on the architecture of transformers, pre-trained models, such as GPT, BERT, etc. [50], are built to tackle NLP tasks. The fine-tuning models achieved exciting results in language understanding and generation. The pre-trained models have also been used in computer vision (CV) tasks. Applying pre-trained ResNet [8] as the backbone, many CV tasks have advanced quickly, such as classification, object detection, segmentation, etc. Since the pre-trained model has advanced at a breakneck pace in NLP and CV, many researchers have begun to study vision-and-language (V-L) learning to improve the downstream task of multimodal learning. Su et al. [51] proposed Visual-Linguistic BERT to adopt the Transformer as the backbone, using visual and linguistic embedding features as inputs to train the model. Tan and Bansal [52] introduced LXMERT, which not only contains a visual and textual encoder but also contains cross-modality encoder to build the relationship between images and sentences. Radford et al. [53] leverage large-scale image–text pairs to build the CLIP model, which jointly trained a visual encoder and textual encoder to predict the correct pairing. Singh et al. [54] proposed FLAVA, which first adopted a dual encoder to obtain unimodal vision and language representation as well as multimodal representation. Yu et al. [55] presented a contrastive captioner, which employed contrastive loss and captioning loss for an image–text encoder–decoder and achieved the state-of-the-art performance with a broad range of downstream tasks.

Even though there is lots of research for multimodal sentiment analysis, few studies have applied pre-trained multimodal representations to this task. Considering the relationship between image and text on social media, this paper utilized the vision–language pre-trained model and the unimodal pre-trained method to build the network for the sentiment analysis task. Specifically, vision–language representation is applied to extract global fusion features; unimodal representation is employed to construct a fine-grained correlation between image regions and words.

3. Proposed Model

3.1. Overview

Text and image data in social media often exist at the same time. However, sometimes, a single text may correspond to multiple images, which is more complicated. Considering that

text and image information are equally important, this paper mainly focuses on the social data that a text corresponds to an image, and the image–text multimodal sentiment classification task is defined as follows. Given image–text pairs $P = \{(I_1, T_1), \ldots (I_i, T_i), \ldots (I_n, T_n)\}$ and the corresponding label set of the image–text pairs $L = \{l_1, \ldots l_i, \ldots l_n\}$, n is the total number of pairs in the set. For an image–text pair (I_i, T_i), I_i denotes a single image, and T_i denotes the corresponding texts. The sentiment label is $l_i \in$ {positive, negative, neutral}. The goal of the multimodal method is to predict the sentimental polarity correctly. The framework of the proposed model is shown in Figure 2. GLFN is composed of three modules: global feature learning, local fusion learning, and an integration network. Global fusion learning first extracts the overall fusion features from the pre-trained image and text representation with an attention mechanism. Then, fine-grained local fusion learning explores the detailed correlation between image content and words. Finally, an integration network is applied to aggregate the sentimental information from global fusion features and local correlation features to conduct the result.

3.2. Global Fusion Learning

The representations of image and text are from pre-trained vision–language CLIP [54]. Since the model used abundantly text–image pair data from the internet for training, natural language is referenced in the visual concepts that can be learned in this process. That is, visual representations can be learned from natural language information, and likewise, textual representations somehow represent image information. We argue that a pretrained multimodal approach can obtain shallow global fused features of image and text, considering the relationship between visual and textual representations. Multidimensional features are used to represent global visual content I_G and textual content T_G to retain more features:

$$I_G = CLIP_{img}(I) \tag{1}$$

$$T_G = CLIP_{txt}(T). \tag{2}$$

The attention mechanism is applied to extract adequate information from shallow global visual and textual fusion features. The attention operation first involves a scale-dot product of query Q and keys K, and the softmax function is used to build a weight map to measure the critical parts in fusion features. The output can be calculated by the weighted sum of values V. Query, keys, and values are the linear projection of the input feature $M = \{m_1, m_2, \ldots, m_{\mathrm{n}}\} \in \mathbb{R}^{d_m \times \mathrm{n}}$. To achieve better performance, multi-head attention is employed in this paper; when there is h head in the attention, the multi-head attention mechanism can be expressed as follows:

$$Q = W_q M,\ K = W_k M,\ V = W_v M \tag{3}$$

where $W_q,\ W_k,\ W_v \in \mathbb{R}^{d_m \times d_m}$ are parameters that need to be learned during training. For each head i, the output O_i is calculated as follows:

$$O_i = softmax\left(Q_i {K_i}^T / \sqrt{d_h}\right) V_i, \tag{4}$$

where $d_h = d_m/h$, $Q_i,\ K_i,\ V_i \in \mathbb{R}^{d_h \times n}$ are the split parts from Q, K, and V, and there are h split parts of keys, query, and value for attention calculation. The output of multi-head attention is shown as follows:

$$O = W_o concat(O_1, O_2, \ldots, O_h). \tag{5}$$

The parameter $W_o \in \mathbb{R}^{d_m \times d_m}$ and output feature O keep the same dimensions as the input features.

So, for given image fusion features I_G and text fusion features T_G, the global image fusion features F_{IG} and the global text fusion features F_{TG} can be represented as follows:

$$O_{Ii} = Softmax\left(\frac{Q_{Ii}{K_{Ii}}^T}{\sqrt{d_{Ih}}}\right)V_i, \tag{6}$$

$$F_{IG} = W_{Io}concat(O_{I1}, O_{I2}, \ldots, O_{Ih}), \tag{7}$$

$$O_{Ti} = Softmax\left(\frac{Q_{Ti}{K_{Ti}}^T}{\sqrt{d_{Th}}}\right)V_i, \tag{8}$$

$$F_{TG} = W_{To}Concat(O_{T1}, O_{T2}, \ldots, O_{Th}), \tag{9}$$

where $d_{Ih} = d_{Im}/Ih$, $d_{Th} = d_{Tm}/Th$, and Ih and Th are the head number for image-based attention and text-based attention, respectively.

3.3. Local Fusion Learning

There are two modules in local fusion learning: a coarse cross-modal module and a fine-grained interaction module. A coarse cross-modal module builds rough interactions between image content and sentences. The fine-grained interaction module further explores the correlation between image region and words; the structure of local fusion learning is shown in Figure 3.

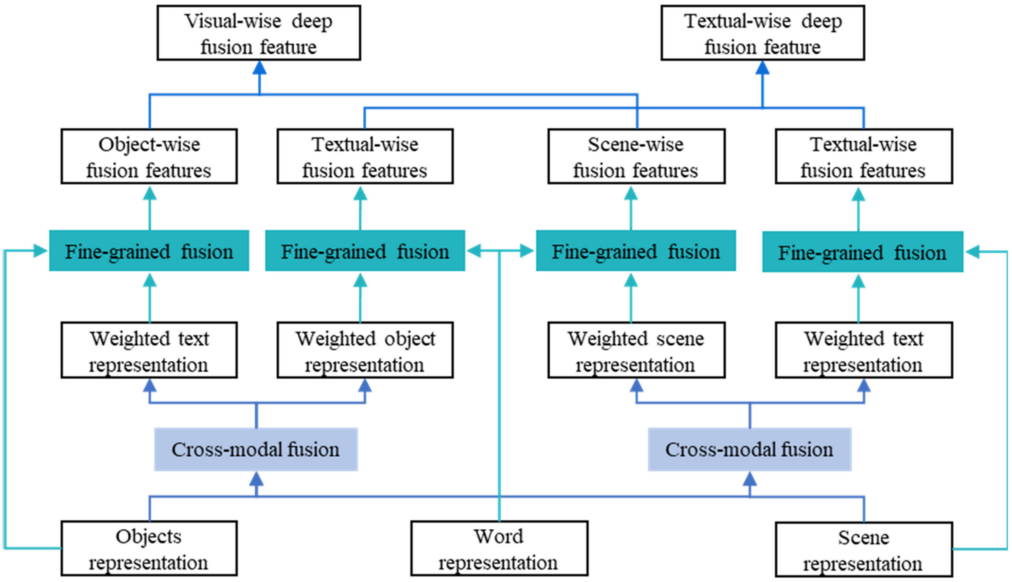


Figure 3. The framework of local fusion learning.

3.3.1. Cross-Modal Module

For the input image, we consider that the image content information is the combination of scenes and objects, the importance of objects and scenes is equal, and this information can be extracted from the pretrained model. Objects' representations and their corresponding regions can be obtained by the Faster RCNN [56], which is trained on the Visual Genomes dataset [11]. Top k region proposals are selected for each image as image object content representations. We detect scene representations by utilizing the VGG model, which is trained on the places 365 dataset [12]. For the input sentence, we apply GloVe [39] to represent each word, and a bidirectional LSTM is employed to summarize the context information of the sentence. The process can be expressed as follows:

$$R_{o'} = Faster\ RCNN(I), \tag{10}$$

$$R_{s'} = VGG(I), \tag{11}$$

$$R_{w'} = BiLSTM(GloVe(T)). \tag{12}$$

To project the representations into the same d dimension space to calculate the weight map, linear projection is applied:

$$R_o = W_o R_{o'} + B_o, \tag{13}$$

$$R_s = W_s R_{s'} + B_s, \tag{14}$$

$$R_w = W_w R_{w'} + B_w, \tag{15}$$

where $R_o \in \mathbb{R}^{\mathrm{d}\times k}$, $R_s \in \mathbb{R}^{\mathrm{d}\times 1}$, $R_w \in \mathbb{R}^{\mathrm{d}\times \mathrm{c}}$, and c is the number of words in the sentence. For the given object features R_o and word features R_w, the object word fusion matrix can be computed as:

$$W_{ow} = {R_o}^T R_w. \tag{16}$$

The fusion matrix $W_{ow} \in \mathbb{R}^{\mathrm{k}\times c}$ represents the affinity between objects from the image and words in a sentence. Specifically, W_{woij} represents the affinity between the ith image region and the jth word. The attention weights between image objects and words can be obtained by softmax function. We obtain word-specific image object region representation $\overline{A_o}$ and region-specific word representation $\overline{A_{ow}}$ by combining the representations and weight map:

$$A_o = softmax\left(\frac{{W_{ow}}^T}{\sqrt{d}}\right), \overline{A_o} = A_o\ {R_o}^T. \tag{17}$$

$$A_{ow} = softmax\left(\frac{W_{ow}}{\sqrt{d}}\right), \overline{A_{ow}} = A_{ow}{\mathrm{R}_w}^T. \tag{18}$$

In a similar way, by replacing R_o with R_s in the Equation (16), W_{sw} can be computed, which shows the affinity between scene and words. When using W_{sw} and R_s and R_w, such as in Equations (17) and (18), we can calculate word-specific image scene representation $\overline{A_s}$ and scene-specific word representation $\overline{A_{sw}}$:

$$W_{sw} = {R_s}^T R_w, \tag{19}$$

$$A_s = softmax\left(\frac{{W_{sw}}^T}{\sqrt{d}}\right), \overline{A_s} = A_s\ {R_s}^T. \tag{20}$$

$$A_{sw} = softmax\left(\frac{W_{sw}}{\sqrt{d}}\right), \overline{A_{sw}} = A_{sw}{\mathrm{R}_w}^T. \tag{21}$$

3.3.2. Fine-Grained Fusion Module

Although the cross-modal module builds the correlation between image objects and words, it is difficult to perfectly align words and image regions because the relationship between images and sentences is complex in social media. Therefore, a fine-grained fusion module is applied to control the coarse fusion feature adaptively, aiming to eliminate the noises generated in the cross-modal module and explore deeper interaction between image region and words. To the extent of matched parts fusion features and suppressing the mismatched pairs of fusion features, in the beginning, a gate is designed to compute the matching value of each fusion feature. Then, the gate weights control how much information is useful to the fusion cross-modal features. The gate weight will be high when the word and image content match well. On the contrary, when the image content and word are unpaired, the weight is low, and the gate weights in the visual aspect are shown as follows:

$$\mathrm{G}_{wo} = \mathrm{s}igmoid\left(sum\left(R_o \odot \overline{A_{ow}}^T\right)\right), \tag{22}$$

$$\mathrm{G}_{ws} = \mathrm{s}igmoid\left(\mathrm{sum}\left(R_s \odot \overline{A_{sw}}^T\right)\right), \tag{23}$$

where $\odot$ denotes the element-wise product, $G_{wo} \in \mathbb{R}^{1\times k}$ is the text-object region matching weights in the visual aspect, and $G_{ws} \in \mathbb{R}^{1\times 1}$ is the text-scene matching weights in the visual aspect. We can obtain the fine-grained fusion features by the gate weights and fused modality-based fusion features. To preserve the modality-based information which is not intensively fused, we further integrate the fine-grained visual fusion feature with original features to obtain the object wise fine-grained text-referred fusion O_{wo} and scene-wise fine-grained text-referred fusion O_{ws}.

$$O_{wo} = MLP\left(G_{wo} \odot concat\left(R_o, \overline{A_{ow}}^T\right)\right) + R_o, \tag{24}$$

$$O_{ws} = MLP\left(G_{ws} \odot concat\left(R_s, \overline{A_{sw}}^T\right)\right) + R_s. \tag{25}$$

MLP is a two-layer perceptron operation.

Symmetrically, we can calculate the weight map, and textual-wise fine-grained object-referred features O_{ow} and textual-wise fine-grained scene-referred features O_{sw}:

$$G_{ow} = sigmoid(sum(\overline{A_o}^T \odot R_w)), \tag{26}$$

$$G_{sw} = sigmoid\left(sum\left(\overline{A_s}^T \odot R_w\right)\right), \tag{27}$$

$$O_{ow} = MLP\left(G_{ow} \odot concat\left(R_w, \overline{A_o}^T\right)\right) + R_w, \tag{28}$$

$$O_{sw} = MLP\left(G_{sw} \odot concat\left(R_w, \overline{A_s}^T\right)\right) + R_w, \tag{29}$$

where $G_{ow} \in \mathbb{R}^{1\times c}$, $G_{sw} \in \mathbb{R}^{1\times c}$. We finally aggregate different fine-grained fusion features in spatial dimension based on the modality and obtain visual-wise local fusion features F_{IL} and textual-wise local fusion features F_{TL}:

$$F_{IL} = concat(O_{wo}, O_{ws}), \tag{30}$$

$$F_{TL} = concat(O_{ow}, O_{sw}). \tag{31}$$

3.4. Integration Network

The network aims to integrate the global and local fusion features effectively and extract valuable information for sentiment analysis. Considering the architecture of global fusion learning and local fusion learning, we adopt different methods to deal with the features. For global fused parts, since the attention mechanism has been applied to explore the deeper effective information for fusion, we apply one-dimensional max-pooling to the global fusion features and on a spatial dimension, producing the features related to sentiment:

$$O_{IG} = maxpooling(F_{IG}), \tag{32}$$

$$O_{TG} = maxpooling(F_{TG}). \tag{33}$$

We employ the attention mechanism on the spatial dimension for the local fused features to aggregate the local image features F_{IL} and local text features F_{TL} with the built spatial-wise attention weight map:

$$att_I = Softmax(MLP(F_{IL})), \tag{34}$$

$$att_T = Softmax(MLP(F_{TL})), \tag{35}$$

where $att_I \in \mathbb{R}^{1\times k}$, $att_T \in \mathbb{R}^{1\times c}$. The final local fusion features can be expressed as follows:

$$O_{IL} = att_I {F_{IL}}^T, \tag{36}$$

$$O_{TL} = att_T {F_{TL}}^T. \tag{37}$$

After obtaining the final global and local fused representations, we concatenate them as the input of a two-layer perceptron operation for the sentiment classification task:

$$\hat{l} = softmax(MLP(concat(O_{IG}, O_{TG}, O_{IL}, O_{TL}))). \quad (38)$$

The model is trained by minimizing the cross-entropy loss with the Adam operation:

$$\mathcal{L} = \sum l_i log\hat{l}_i. \quad (39)$$

4. Experiments

This section describes the experimental results of the proposed model on two open datasets. This section consists of four parts: experimental data and model setup, baseline and comparison, ablation studies, and visualization.

4.1. Dataset and Setup

Niu et al. [26] established the public multimodal sentiment analysis datasets, which were collected from Twitter, including MVSA-Single and MVSA-Multiple. There are 5129 image–text pairs in MVSA-Single. An annotator labeled the image and text with one sentiment polarity form as positive, neutral, and negative, respectively. MVSA-Multiple contains 196,00 image–text pairs. Three annotators give the label to the image and text independently, and the judgment of each annotator is not affected by others. For a fair comparison, we preprocess the two datasets following the previous method [26,42,43]: first, remove the pairs which have different labels between image and text; when one label is positive(negative), and the other corresponding content is neutral, the label of this pair is regarded as positive(negative). As a result, we obtain the new MVSA-Single, which has 4511 image–text pairs and MVSA-Multiple datasets with 17,024 image–text pairs for the experiment.

In this study, the training, validation, and test sets are split with a ratio of 8:1:1. We utilize Adam as the optimizer method. The initial learning rate is 5×10^{-5}, and an exponential decay is applied with gamma equal to 0.5 for every five epochs. The batch size of MVSA-Single is 32, and the batch size of MVSA-Multiple is 128.

The maximum word number for each sentence is 50, the head number of the global fusion attention mechanism is 8, and we select the top 3 image region features as local image presentation. The dimension for projecting the local features into the same space is 512. The framework of the model is implemented by PyTorch.

4.2. Baseline and Comparison

We list six studies utilizing the deep learning method for MVSA dataset as follows:

1. MultiSentiNet [13]: employed CNN to obtain objects and scene deep semantic features of the image and utilized visual feature guided attention LSTM to extract important word features; all these features are aggregated for the sentiment analysis.
2. Co-Memory [42]: proposed an iterative co-memory model with an attention mechanism by considering the relationship between image and text; the network explored the interaction between visual and textual information multiple times to analyze users' sentiment.
3. FENet [43]: introduced an interactive information fusion mechanism, which learned the deep modality-specific fusion representation and built an information extraction network to extract information more effectively for the multimodal sentiment task.
4. MVAN [45]: proposed a multi-view attention-based interaction model, which built an iterative scene–text co-memory network, as well as an iterative object–text co-memory network to obtain semantic image–text features for the sentiment analysis.
5. CLMLF [46]: introduced contrastive learning with a multi-layer transformer-based fusion method. Two contrastive learning tasks, label-based contrastive learning and

data-based contrastive learning, are proposed for training to help the model learn the common features for sentiment analysis.

6. ITIN [47]: developed an image–text interaction network to align the information between image region and words and preserved the valid region–word pairs with a cross-modal gating module for effective fusion features. The unimodal features are combined with cross-modal fusion features for the sentiment classification.

Same as the previous studies, we use the accuracy rate and $F1$-score as the experimental evaluation metrics. The calculation is as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP}, \\ Recall &= \frac{TP}{TP+FN}, \\ F1 &= \frac{2\times Precision\times Recall}{Precision+Recall}, \\ Accuracy &= \frac{TP+TN}{TP+TN+FP+FN}. \end{aligned} \tag{40}$$

TP is the number of samples that obtain the correct prediction with the positive label, and FP is the number of samples that wrongly indicate the prediction as positive while the label is negative. TN denotes the number of samples that are marked as negative correctly. FN presents the number with a negative prediction and positive label.

The experimental results comparisons between the proposed model and the baselines are shown in Table 1. The results of baseline methods were retrieved from published papers.

Table 1. Comparative results of GLFN and baselines on MVSA datasets, ACC represents accuracy and F1 denotes F1 score (↑ denotes the value is larger, the performance of model is better).

Model	MVSA-Single		MVSA-Multiple	
	ACC↑	**F1↑**	**ACC↑**	**F1↑**
MultiSentiNet	0.6984	0.6963	0.6886	0.6811
Co-Memory	0.7051	0.7001	0.6892	0.6883
FENet-GloVe	0.7254	0.7232	0.7057	0.7038
MVAN	0.7298	0.7298	0.7236	0.7230
CLMLF	0.7533	0.7346	0.7200	0.6983
ITIN	0.7519	0.7497	0.7352	0.7349
Proposed model (GLFN)	**0.7721**	**0.7642**	**0.7587**	**0.7399**

The results show that the MultiSentiNet obtained the worst performance among the models; even though the model considered the influence of visual information on the text, it ignored the influence of textual information on the image, and the interaction between image and text is shallow. Co-Memory considered the mutual influence and proposed an interactive fusion module to analyze the sentiment and is better than MultiSentiNet, but the coarse-grained attention mechanism may cause information loss and still need to be improved. FENet applied symmetry interactive information fusion module and the information extraction module to obtain informative representations, which separated the analysis into two sections, making the model more effective than the previous model. Since MVAN proposed a multi-view attention model to build the correlation between objects/scenes and words and integrated the features to predict results, it is slightly superior to FENet. CLMLF applied label-based and data-based contrastive learning for the sentiment task, which learned more features related to sentiment and is better than MVAN. ITIN model considered the relationship between image region and words and built deep region information based fusion features for the classification, achieving the best results among the baseline. Our model is competitive with other baseline models on the MVSA dataset. The model extracts the coarse global fusion features to obtain whole fusion information for image–text pairs and explores fine-grained local fusion features for detailed correlation between corresponding regions from images and words in sentences, which utilizes the

interaction comprehensively for multimodal sentiment analysis. As a result, for the MVSA-Single dataset, the proposed model outperforms the existing base model ITIN and CLMLF in terms of accuracy and F1 score with 2.02% and 1.45%. For the MVSA-Multiple dataset, the model improved by 2.35% in accuracy and is superior to other models in $F1$ score.

4.3. Ablation Studies

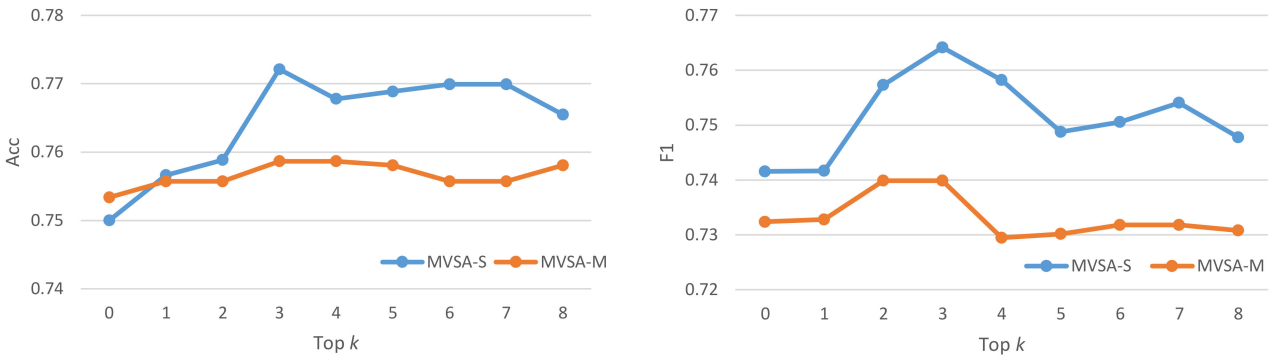
To verify the effectiveness of each proposed module of GLFN, we carry out ablation experiments on the MVSA dataset. We remove textual global fusion learning (wo-TGF), visual global fusion learning (wo-IGF), textual local fusion learning (wo-TLF), and visual local fusion learning (wo-ILF) of GLFN to evaluate the influence of each part. In addition, we further investigate the importance of the cross-modal fusion module and fine-grained fusion module of local fusion learning, removing the cross-modal fusion module (wo-CM) and fine-grained fusion module (wo-FF) independently. The results of the ablation experiments are reported in Table 2.

Table 2. Ablation study results of the GLFN model (↑ denotes the value is larger, the performance of model is better).

Model	MVSA-Single		MVSA-Multiple	
	ACC↑	**F1↑**	**ACC↑**	**F1↑**
wo-TGF	0.7442	0.7366	0.7398	0.7189
wo-IGF	0.7597	0.7497	0.7403	0.7268
wo-TLF	0.7637	0.7517	0.7396	0.7218
wo-ILF	0.7597	0.7497	0.7430	0.7281
wo-CM	0.7580	0.7522	0.7428	0.7253
wo-FF	0.7641	0.7558	0.7437	0.7293
GLFN	0.7721	0.7642	0.7587	0.7399

From the results, we can observe that the full version of GLFN achieves the best results; removal of either part of the model will affect the performance. This indicates that all these parts are adequate for the sentiment task. For global fusion learning, textual fusion features are more important than visual features for the prediction. The importance of the sentiment of visual and textual local fusion features is different by the datasets, which strongly depend on the relationship between images and texts. The additional investigation for local fusion shows that the cross-modal fusion module is more important than the fine-grained fusion module, and the fine-grained fusion further enhances the performance of the model. Compared with other parts, textual global fusion features are more effective for the prediction.

We also consider the number of region proposals in the image as a hyperparameter that affects the interaction between image and text, and the performance of sentiment analysis. In addition, we conduct experiments with different top k region proposals where objects exist. Figure 4a,b show the experimental results of accuracy and F1-score. MVSA-S represents the results obtained from MVSA-Single, and MVSA-M denotes the results performed on MVSA-Multiple. The figures show that when $k = 3$, both the accuracy and F1-score reach the maximum values on the two datasets. The values show that correlation discovery between image regions and words can be found within three region proposals in the image, and the local learning is more valid when the number of region proposals equals three. Therefore, for all the experiments reported in this paper, the region number is set to three. Here, $k = 0$ denotes the local fusion learning is removed from the model, which has a significant influence on the performance. The results indicate the effectiveness of local fusion learning.

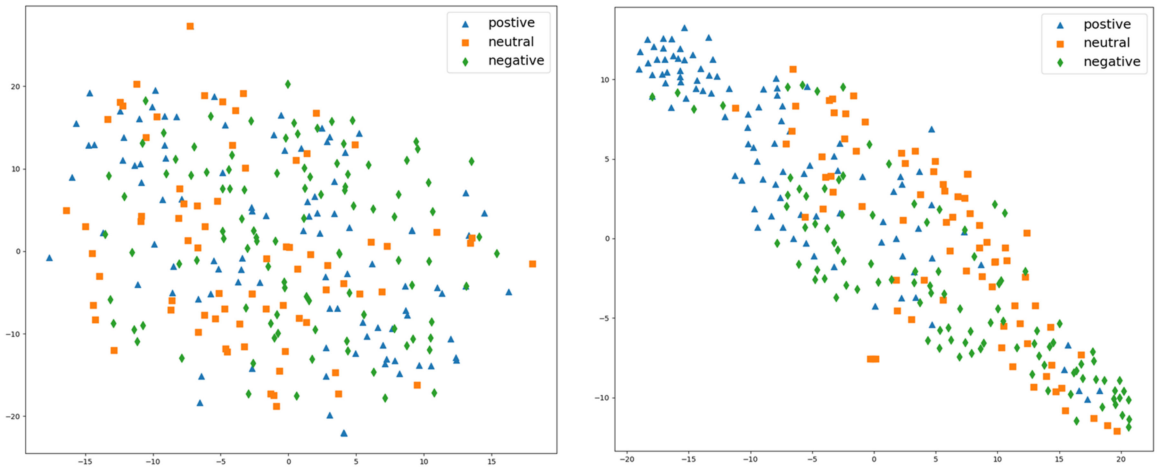


(**a**) (**b**)

Figure 4. The experimental results for different values of hyperparameter k. (**a**) The accuracy of GLFN with different top k. (**b**) The F1-score of GLFN with different top k.

4.4. Visualization

We conduct a visualization experiment on parts of the MVSA-Single dataset to demonstrate the effectiveness of global fusion learning. The image's global fusion and initial representation are visualized by dimensionality reduction. We use the TSNE algorithm to reduce the dimensions of the visual features and project the representations into a two-dimensional space to visualize them, as shown in Figure 5. Figure 5a is the visualization of original visual features, and Figure 5b is the visualization of attention-based visual fusion features. The three marking symbols in the figure represent three different labels of sentiment: positive, neutral, and negative. From the figure, we can see that after the attention mechanism of fusion learning, the distribution distance between the different categories is more prominent. That is, the intra-class distance decreases, and the spacing between classes increases. The figure shows that global fusion learning has the ability to distinguish the valid information that helps improve the performance of sentiment analysis.



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

Figure 5. Visualization results of visual features projected to two-dimensional space. (**a**) Visualization of the initial visual features. (**b**) Visualization of the attention-based features.

We visualize the attention weight of the words and image regions of the local fusion learning part. Figure 6 illustrates two examples of visualization. The column shows region and scene representations with the green weight color, and the bottom row represents the words with red attention weight color. A darker color represents greater attention and vice versa. We can see that the contribution of image regions and words is different in Figure 6a. The area of the sky has the most contribution to the sentiment analysis, and the

words 'Gorgeous Milky Way' are more effective than others for the prediction. In addition, the road region has little attention in this image–text pair since both content and emotion correlate less with that region. In Figure 6b, the region of the cars has the most significant contribution, the area of the scene gives little contribution to the sentiment analysis, and the corresponding text is a little complex; the word 'disgraceful' and phrase 'nearly run people off' are the crucial parts for the analysis.

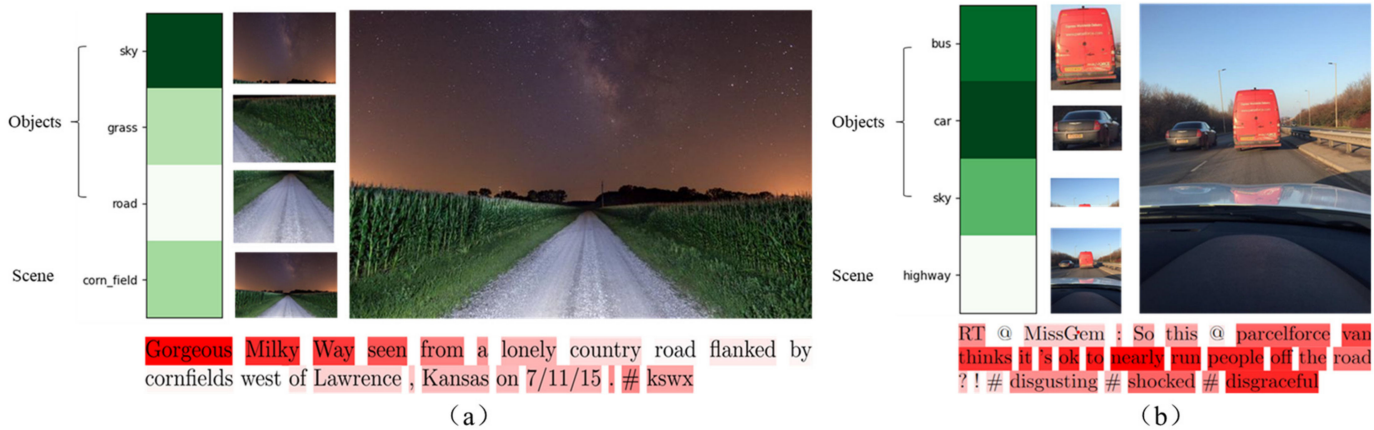


Figure 6. Visualization of the contribution of image region features and textual features for multimodal sentiment analysis. (**a**) expresses a positive feeling, (**b**) shows a negative mood.

5. Conclusions

Social media multimodal sentiment analysis is a challenging task. We have proposed a global local fusion neural network (GLFN) for the sentiment prediction task. The model considers the relationship between image and text, combining the general fusion information extracted from global fusion learning and local fine-grained fusion information obtained from local fusion learning to explore essential features related to the sentiment. To be specific, the pre-trained vision–language model is employed as the input of global fusion learning to obtain the comprehensive overall fusion features. In local fusion learning, scene and object representations construct a deep correlation to words as fine-grained fusion features for specified relation discovery between image and sentence. The integration network aggregates visual and textual context features as integrated fusion information for effective sentiment prediction. Experiment results and the comparison demonstrate that our model significantly improves the sentiment classification performance in the multimodal dataset. Even though we obtained promising results, there is a limitation to the model. Since there are some posts in which image and text are unrelated, this leads the sentiment expression to rely on the independent features, which may limit the performance of GLFN. So, in future work, we plan to improve the integration network by exploring the complicated relationship between images and texts to build more effective features by considering the ratio of fused information and independent information on the effect of sentiment. Furthermore, we want to extend the model to multimodal aspect sentiment analysis since the local fusion learning can align the image and text information and explore the detailed correlation between the corresponding pairs.

Author Contributions: Conceptualization, X.H. and M.Y.; methodology, X.H.; software, X.H.; validation, X.H. and M.Y.; formal analysis, X.H.; investigation, X.H.; writing—original draft preparation, X.H.; writing—review and editing, M.Y.; visualization, X.H.; supervision, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are from public datasets that can be downloaded from https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/ (accessed on 1 January 2016).

Conflicts of Interest: The authors declare no conflict of interest regarding the publication of this paper.

References

1. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; Pantic, M. A Survey of Multimodal Sentiment Analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
2. Chakraborty, K.; Bhattacharyya, S.; Bag, R. A Survey of Sentiment Analysis from Social Media Data. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 450–464. [CrossRef]
3. Shao, X.; Tang, G.; Bao, B.-K. Personalized Travel Recommendation Based on Sentiment-Aware Multimodal Topic Model. *IEEE Access* **2019**, *7*, 113043–113052. [CrossRef]
4. Anwar Hridoy, S.A.; Ekram, M.T.; Islam, M.S.; Ahmed, F.; Rahman, R.M. Localized Twitter Opinion Mining Using Sentiment Analysis. *Decis. Anal.* **2015**, *2*, 8. [CrossRef]
5. Kauffmann, E.; Peral, J.; Gil, D.; Ferrández, A.; Sellers, R.; Mora, H. Managing Marketing Decision-Making with Sentiment Analysis: An Evaluation of the Main Product Features Using Text Data Mining. *Sustainability* **2019**, *11*, 4235. [CrossRef]
6. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College PA USA, New York, NY, USA, 13 October 2004; Association for Computing Machinery: New York, NY, USA; pp. 205–211.
7. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Bengio, Y., LeCun, Y., Eds.; Conference Track Proceedings. 2015.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society, Las Vegas, NV, USA, 1 June 2016; pp. 770–778.
9. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
10. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Santiago, Chile, 2015; pp. 19–27.
11. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
12. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1452–1464. [CrossRef]
13. Xu, N.; Mao, W. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6 November 2017; Association for Computing Machinery: New York, NY, USA; pp. 2399–2402.
14. Kim, J.; Koh, J.; Kim, Y.; Choi, J.; Hwang, Y.; Choi, J.W. Robust Deep Multi-Modal Learning Based on Gated Information Fusion Network. In Proceedings of the Computer Vision—ACCV 2018, Perth, Australia, 2–6 December 2018; Jawahar, C.V., Li, H., Mori, G., Schindler, K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 90–106.
15. Zhou, Y.; Yang, Y.; Liu, H.; Liu, X.; Savage, N. Deep Learning Based Fusion Approach for Hate Speech Detection. *IEEE Access* **2020**, *8*, 128923–128929. [CrossRef]
16. Akhtar, M.S.; Chauhan, D.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-Task Learning for Multi-Modal Emotion Recognition and Sentiment Analysis. 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), Minneapolis, MN, USA, 2–7 June 2109; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 370–379.
17. Han, W.; Chen, H.; Poria, S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7–11 December 2021; pp. 9180–9192.
18. 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, New York, NY, USA, 12–16 October 2020; Association for Computing Machinery: New York, NY, USA; pp. 1122–1131, ISBN 978-1-4503-7988-5.
19. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.; Poria, S. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 19–22 October 2021; pp. 6–15.
20. 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]

21. 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), Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2236–2246.
22. Kruk, J.; Lubin, J.; Sikka, K.; Lin, X.; Jurafsky, D.; Divakaran, A. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. 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; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 4622–4632.
23. Vempala, A.; Preoţiuc-Pietro, D. Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 2830–2840.
24. Henning, C.A.; Ewerth, R. Estimating the Information Gap between Textual and Visual Representations. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, 6–9 June 2019; Association for Computing Machinery: New York, NY, USA, 2017; pp. 14–22.
25. Otto, C.; Springstein, M.; Anand, A.; Ewerth, R. Characterization and Classification of Semantic Image-Text Relations. *Int. J. Multimed. Info. Retr.* **2020**, *9*, 31–45. [CrossRef]
26. Niu, T.; Zhu, S.; Pang, L.; El Saddik, A. Sentiment Analysis on Multi-View Social Data. In Proceedings of the MultiMedia Modeling, Miami, FL, USA, 4–6 January 2016; Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 15–27.
27. Park, S.; Kim, Y. Building Thesaurus Lexicon Using Dictionary-Based Approach for Sentiment Classification. In Proceedings of the 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, USA, 8–10 June 2016; pp. 39–44.
28. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Barcelona, Spain, 2002; pp. 79–86.
29. Borg, A.; Boldt, M. Using VADER Sentiment and SVM for Predicting Customer Response Sentiment. *Expert Syst. Appl.* **2020**, *162*, 113746. [CrossRef]
30. Basiri, M.E.; Nemati, S.; Abdar, M.; Cambria, E.; Acharya, U.R. ABCDM: An Attention-Based Bidirectional CNN-RNN Deep Model for Sentiment Analysis. *Future Gener. Comput. Syst.* **2021**, *115*, 279–294. [CrossRef]
31. Siersdorfer, S.; Minack, E.; Deng, F.; Hare, J. Analyzing and Predicting Sentiment of Images on the Social Web. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; Association for Computing Machinery: New York, NY, USA; pp. 715–718.
32. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.-F. Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 223–232.
33. Yang, Y.; Jia, J.; Zhang, S.; Wu, B.; Chen, Q.; Li, J.; Xing, C.; Tang, J. How Do Your Friends on Social Media Disclose Your Emotions? In Proceedings of the AAAI Conference on Artificial Intelligence, Québec, QC, Canada, 27–31 July 2014; Volume 28. [CrossRef]
34. Chen, T.; Borth, D.; Darrell, T.; Chang, S.-F. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. *arXiv* **2014**, arXiv:1410.8586.
35. Song, K.; Yao, T.; Ling, Q.; Mei, T. Boosting Image Sentiment Analysis with Visual Attention. *Neurocomputing* **2018**, *312*, 218–228. [CrossRef]
36. Wu, L.; Qi, M.; Jian, M.; Zhang, H. Visual Sentiment Analysis by Combining Global and Local Information. *Neural Process Lett.* **2020**, *51*, 2063–2075. [CrossRef]
37. Baecchi, C.; Uricchio, T.; Bertini, M.; Del Bimbo, A. A Multimodal Feature Learning Approach for Sentiment Analysis of Social Network Multimedia. *Multimed. Tools Appl.* **2016**, *75*, 2507–2525. [CrossRef]
38. Hu, A.; Flaxman, S. Multimodal Sentiment Analysis to Explore the Structure of Emotions. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19 July 2018; pp. 350–358.
39. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543.
40. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 October 2016; pp. 439–448.
41. 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; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 1103–1114.
42. Xu, N.; Mao, W.; Chen, G. A Co-Memory Network for Multimodal Sentiment Analysis. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 929–932.

43. Jiang, T.; Wang, J.; Liu, Z.; Ling, Y. Fusion-Extraction Network for Multimodal Sentiment Analysis. In Proceedings of the Advances in Knowledge Discovery and Data Mining, Singapore, 11–14 May 2020; Lauw, H.W., Wong, R.C.-W., Ntoulas, A., Lim, E.-P., Ng, S.-K., Pan, S.J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 785–797.
44. Hu, X.; Yamamura, M. Two-Stage Attention-Based Fusion Neural Network for Image-Text Sentiment Classification. In Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing, Singapore, 18–20 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–7.
45. Yang, X.; Feng, S.; Wang, D.; Zhang, Y. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Trans. Multimed.* **2021**, *23*, 4014–4026. [CrossRef]
46. Li, Z.; Xu, B.; Zhu, C.; Zhao, T. CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. *arXiv* **2022**, arXiv:2204.05515.
47. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal Sentiment Analysis with Image-Text Interaction Network. *IEEE Trans. Multimed.* **2022**, 1. [CrossRef]
48. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, 2–4 May 2013; Workshop Track Proceedings. Bengio, Y., LeCun, Y., Eds.;
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
50. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-Trained Models for Natural Language Processing: A Survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
51. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-Training of Generic Visual-Linguistic Representations. *arXiv* **2019**, arXiv:1908.08530.
52. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 5100–5111.
53. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning PMLR, San Francisco, CA, USA, 1 July 2021; pp. 8748–8763.
54. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. FLAVA: A Foundational Language and Vision Alignment Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
55. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners Are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
56. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 28.