*mathematics*

MDPI

*Article*

# A Semantic Enhancement Framework for Multimodal Sarcasm Detection

**Weiyu Zhong †, Zhengxuan Zhang †, Qiaofeng Wu, Yun Xue and Qianhua Cai ***

School of Electronics and Information Engineering, South China Normal University, Foshan 528225, China; zhong_wy@m.scnu.edu.cn (W.Z.); zx_zhang@m.scnu.edu.cn (Z.Z.); scnu_wqf@m.scnu.edu.cn (Q.W.); xueyun@m.scnu.edu.cn (Y.X.)

* Correspondence: caiqianhua@m.scnu.edu.cn

† These authors contributed equally to this work.

**Abstract:** Sarcasm represents a language form where a discrepancy lies between the literal meanings and implied intention. Sarcasm detection is challenging with unimodal text without clearly understanding the context, based on which multimodal information is introduced to benefit detection. However, current approaches only focus on modeling text–image incongruity at the token level and use the incongruity as the key to detection, ignoring the significance of the overall multimodal features and textual semantics during processing. Moreover, semantic information from other samples with a similar manner of expression also facilitates sarcasm detection. In this work, a semantic enhancement framework is proposed to address image–text congruity by modeling textual and visual information at the multi-scale and multi-span token level. The efficacy of textual semantics in multimodal sarcasm detection is pronounced. Aiming to bridge the cross-modal semantic gap, semantic enhancement is performed by using a multiple contrastive learning strategy. Experiments were conducted on a benchmark dataset. Our model outperforms the latest baseline by 1.87% in terms of the F1-score and 1% in terms of accuracy.

**Keywords:** multimodal sarcasm detection; contrastive learning; graph neural networks; social media

**MSC:** 18C50



**Citation:** Zhong W.; Zhang Z.; Wu Q.; Xue Y.; Cai Q. A Semantic Enhancement Framework for Multimodal Sarcasm Detection. *Mathematics* **2024**, *12*, 317. https://doi.org/10.3390/math12020317

Academic Editors: Jianping Gou, Weihua Ou, Shaoning Zeng and Lan Du

Received: 25 December 2023
Revised: 9 January 2024
Accepted: 12 January 2024
Published: 18 January 2024



## 1. Introduction

Sarcasm refers to satirical or ironic statements where the literal meaning of the text is converse to the authentic intention of the speaker [1,2]. In recent years, sarcastic utterances have become ubiquitous on social media platforms and in daily life. As such, the detection of sarcasm holds great potential in not only understanding the real sentiment of an individual but also mining the extensive conversation contexts in social discussions. There is an ongoing trend whereby sarcasm detection tasks are attracting a great deal of interest [2,3]. Research on this task can find applications in diverse real-world scenarios, including social media monitoring, sentiment analysis in customer reviews, automated content moderation, and image retrieval systems [4].

In addition to texts, social media posts generally involve information of different modalities. Images form one such category, being considered auxiliary to the text in sarcasm detection. According to Figure 1, the sentence "*What a wonderful weather!*" conveys no sarcastic intention via the words, whereas the attached image, which presents a gray sky, expresses the opposite sentiment to the text. We can thereby identify it as a sarcastic sentence. A key factor in sarcasm detection is the "inconsistency" between the image and text, with recent publications focusing on distinguishing the discrepancy between textual and visual information [2,5–7].

What a wonderful weather!

Weather 's lookin amazing today.

**Figure 1.** Examples of sarcastic social media posts.

In general, three processing steps can be distinguished when performing multimodal sarcasm detection: textual and visual information extraction, intra- and inter-modal interaction, and similarity determination and classification. Specifically, the exploitation of images is concerned with the employment of object detectors [8], patches [2,6,7], etc. These approaches are carried out through the matching and interaction of the local visual information with the text, and in this manner, the overall sentiment of the image is neglected. As presented in Figure 1, the gray sky conveys a negative sentiment that is opposite to the sentiment of the sentence, which is indeed a substantial clue for consistency identification.

Despite the use of images, sentence syntax also plays a pivotal role in sarcasm detection, being widely applied to text–image interaction processes. In line with the advances in natural language processing and not just syntactic structure, semantic information also affects the sentiment delivery [9–11]. In the state-of-the-art sarcasm-detecting methods, the analysis of semantics is, however, still limited. Furthermore, people typically convey their opinions with similar semantic expressions. According to the two examples in Figure 1, both sentences literally appreciate the weather but are actually sarcastic. Thus, the semantic information can be enhanced by associating the similarities.

However, in cases where the image–text consistency is taken as the only solution for sarcasm identification, misunderstanding of the textual information can be induced. In Figure 2, no sarcastic expression is given in this case. The sentence delivers a positive sentiment about Argentina winning the World Cup, while the image presents a negative sentiment from the coach of the opposing team. As a result, the semantics of the text is indispensable for sentiment prediction. Aside from that, when seeing the different feature spaces of the image and text, there is a semantic gap between the visual and textual features that also considerably affects the resolution of text–image consistency.

Argentina are finally three-time champions after 36-year wait!

**Figure 2.** An example of a non-sarcastic statement with inconsistent text–image sentiment.

To address the above challenges, we propose a novel semantic enhancement framework (SEF) for multimodal sarcasm detection. Specifically, we first obtain the textual vector and the visual features via BERT and ViT, respectively. Then, we obtain the congruity levels of the text and images across different spans by using cross-modal attention with graph neural networks [2]. Considering the overall information of text and images and the semantic distinction, we construct positive and negative examples of text–image pairs from the same batch, compute the overall text–image similarity, and use contrastive learning to optimize the multimodal representation. In addition, we model the semantic relations

within the text through an Intra-modal GNN, which consists of a multi-headed attention network and a graph neural network. The semantic information is enhanced with the integration of other sentences from the same batch via contrastive learning. To the best of our knowledge, the SEF is the first method that highlights the significance of semantics and verifies its effectiveness in multimodal sarcasm detection.

Our contributions are threefold and are as follows:

- We propose a semantic enhancement framework (SEF) for multimodal sarcasm detection tasks that captures the intra- and inter-modal semantic information of both multiple spans and multiple granularities.
- By using contrastive learning, the semantics from text–image pairs within the same batch is exploited for semantic information enhancement, aiming to bridge the semantic gap between visual and textual modalities.

## 2. Related Work

Different from classical text-based sarcasm detection, multimodal sarcasm detection integrates the information from both the text and the image to predict the sarcasm label. The multimodal sarcasm detection task was first proposed by Schifanella et al. [12], who made predictions based on manually extracted multimodal features. Cai et al. [13] devised a hierarchical fusion network to deal with multimodal information and developed a Twitter-based dataset for multimodal sarcasm detection tasks. Xu et al. [5] worked on cross-modal comparison and semantic relations, based on which a decomposition and relationship network was constructed. Yue et al. [14] proposed a novel model that utilizes ConceptNet for prior knowledge incorporation and employed contrastive learning to enhance multimodal satire detection. Qiao et al. [15] introduced MILNet, a Mutual-enhanced Incongruity Learning Network, for multimodal sarcasm detection, addressing issues related to irrelevant information and incomplete input.

Pan et al. [6] set a foundation for identifying the text–image incongruity in the multimodal sarcasm detection task. Since then, a number of studies have proposed methods that model the relationship between graphs and texts by using graph neural networks (GNNs). Liang et al. [7] constructed a heterogeneous graph structure and used a GNN to capture the intra- and inter-modal relations of the image and text. With the integration of object detection, Liang et al. [8] designed Cross-Modal Graph Convolutional Networks to tackle the matching between the text and objects in an image. Li et al. [2] captured both atomic-level and composition-level incongruity between the image and text via a graph attention network. Wen et al. [16] presented a Dual Incongruity Perceiving (DIP) network for multimodal sarcasm detection, addressing the intrinsic incongruity between the image and text in sarcastic data through channel-wise reweighting and Siamese layers. However, these methods use text–image incongruity as the only inference for sarcasm detection, which can cause an incorrect prediction. By contrast, we mitigate this deficiency by constructing a semantic enhancement framework via a GNN. In our model, the text semantics are dedicatedly analyzed during processing. Moreover, other samples from datasets are exploited for semantic enhancement via contrastive learning, which bridges the semantic gap between multiple modalities.

## 3. Methodology

In this section, the semantic enhancement framework (SEF) for multimodal sarcasm detection is described in detail. The architecture of the SEF is presented in Figure 3. Our model consists of three main components: (1) a feature extractor, which employs pretrained BERT and Vision Transformer (ViT) to, respectively, encode texts and images and thus obtain the hidden representations of both modalities; (2) a cross-modal interaction module, which works on the inter-modal interaction and computes the text–image congruity at multi-span and multi-granularity levels; (3) a semantic enhancement module, which introduces textual semantics via the intra-modal interaction of the text and optimizes the textual representation via supervised contrastive learning.

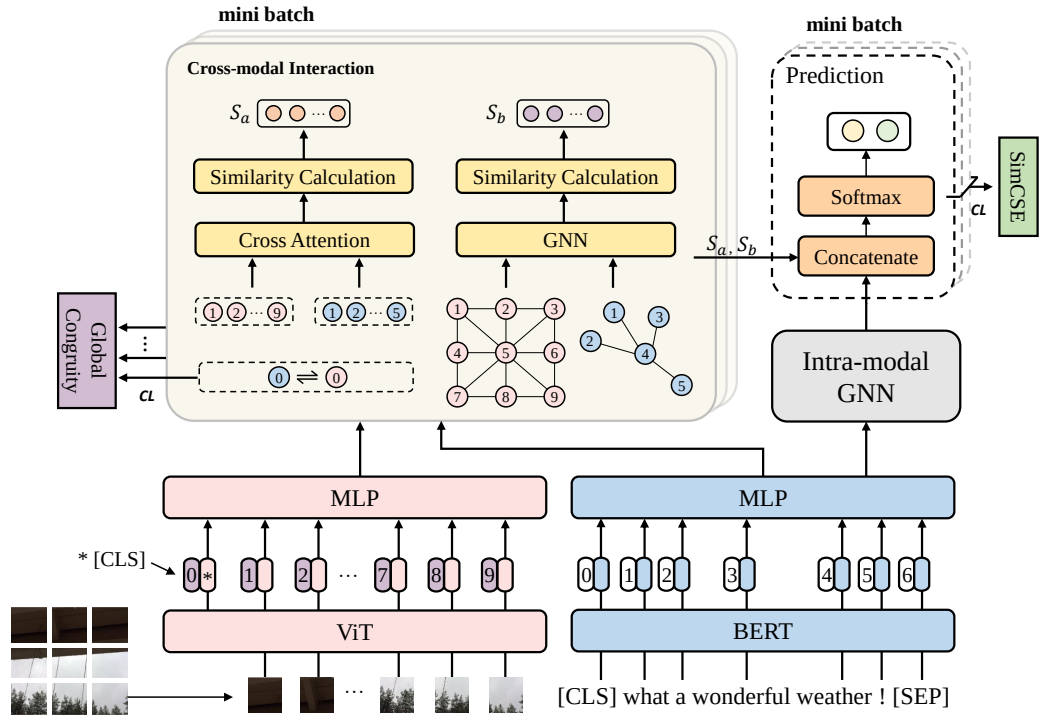


**Figure 3.** Model architecture.

### *3.1. Feature Extractor*

Let $(S_T, S_I)$ be a text–image pair, where $S_T = \{x_1, x_2, \cdots, x_n\}$ is an n-word text and $S_I$ is the attached image. Each word $x_i$ in the text is mapped into a 768-dimensional embedding using the pretrained BERT model [17], $X = BERT([CLS]S_T[SEP])$. We define $X_{cls}$ as the representation of the token $[CLS]$ to denote the sentence information. The textual feature representation is $T$ ($T \in R^{n \times d}$), generated by sending $X$ into a multi-layer perceptron (MLP) to transform it into a 300-dimensional output. On the other hand, assuming the size of image $S_I$ is $L_h \times L_w$, we resize $S_I$ to 224 × 224 pixels [5,7]. The image is then divided into $r$ patches, and these patches are reshaped into a sequence of $Z = \{p_1, p_2, \cdots, p_r\}$. The sequence $Z$ is fed into ViT [18], which is trained on ImageNet for image classification. The output image patch embedding $V = ViT([CLS]Z)$ contains rich visual semantic information. Likewise, the representation $V_{cls}$ of the token $[CLS]$ represents the semantics of the given image $S_I$. Then, the visual feature representation $I$ ($I \in R^{r \times d}$) is also generated via an MLP.

### *3.2. Cross-Modal Interaction*

#### 3.2.1. Token-Level Congruity

As pointed out in the Introduction, text–image congruity is a crucial criterion for multimodal sarcasm detection. Following the idea of Liu et al. [2], we compute the atomic-level congruity and the composition-level congruity between each text token and image patch after inter- and inter-modality fusion, respectively.

**Atomic-level congruity score:** Textual features and visual features are aligned using a multi-headed cross-attention mechanism, based on which the cross-modal features are mapped into the same space:

$$H_i = \text{softmax}\left(\frac{\left(TW_q^i\right)^\top}{\sqrt{d/h}}\left(IW_k^i\right)\right)\left(IW_v^i\right) \tag{1}$$

$$T^* = \text{norm}(T + \text{MLP}([H_1 \parallel H_2 \parallel \cdots \parallel H_h])) \tag{2}$$

where $T \in R^{n\times d}$ and $I \in R^{r\times d}$ are the feature representations of the text and the image; $h$ is the cross-attention head number; $W_q^i, W_k^i, W_v^i \in R^{d\times \frac{d}{h}}$ are the projection matrices of the query, key, and value, respectively; the function "*norm*" stands for the normalization operation; "$\|$" denotes the concatenation operation; $H_i$ is the output of each cross-attention head; and $T^*$ is the textual representation with the alignment of the image. To detect the cross-modal consistency, the consistency score between the $i$-th text token and the $j$-th image patch is computed as follows:

$$G_a = \frac{1}{\sqrt{d}}\left(T^* I^\top\right) \tag{3}$$

Obviously, diversified words have distinguishing effects on sarcasm detection. The atomic-level congruity score $s_a$ is derived via a weighted sum of $G_a$ and the importance score of each token:

$$s_a = \text{softmax}(T^* W_a + b_a)^\top G_a \tag{4}$$

where $W_a \in R^{d\times 1}$ and $b_a \in R^n$ are trainable parameters for the token importance score computation.

**Composition-level congruity score:** At this stage, the textual graph and the visual graph are constructed based on the input text–image pairs. Specifically, in the text graph, each node stands for a text token, and each edge represents the dependency between words, which are extracted by spaCy [2,7]. For the visual graph, each image patch is taken as a graph node, which connects to its adjacent nodes according to the geometrical adjacency in the image. Notably, the graphs of both modalities are undirected and contain self-loops for representation. Subsequently, a graph attention network (GAT) is employed to deal with both textual and visual graphs [19]. By exploiting the masked self-attention layers, the GAT is employed to learn the relative importance between nodes in order to obtain multimodal information on a deeper level. We take the text graph as an example:

$$\alpha_{ij}^l = \frac{\exp\left(\text{LeakyReLU}\left(a_l^\top \left[W_l t_i^l \parallel W_l t_j^l\right]\right)\right)}{\sum_{k\in N_i} \exp\left(\text{LeakyReLU}\left(a_l^\top \left[W_l t_i^l \parallel W_l t_k^l\right]\right)\right)} \tag{5}$$

$$t_i^{l+1} = \alpha_{ii}^l W_l t_i^l + \sum_{j\in N_i} \alpha_{ij}^l W_l t_j^l \tag{6}$$

where $k$, $W_l \in R^{d\times d}$, and $a_l \in R^{2d}$ are learnable parameters of the $l$-layer in the GAT; $\alpha_{ij}^l$ refers to the attention score between node $i$ and its adjacent node $j$; $t_i^l$ is the feature of node $i$ in layer $l$; and $t_i^{l+1}$ is the node output. We define $T^{'} = [t_1^{L^T}, t_2^{L^T}, \cdots, t_n^{L^T}]$ as the textual embedding of the $L^T$-layer in the GAT, involving the complex dependencies of all relevant textual tokens. However, considering the parsing errors and the lack of syntax-related words, the text graph can be unreliable. As such, the weighted sum $b \in R^d$ of each word embedding in $T^*$ is computed, which is further concatenated with $T^{'}$ to demonstrate the congruity:

$$b = \text{softmax}(TW_P + b_P)^\top T^* \tag{7}$$

$$T^{''} = T^{'} \parallel b \tag{8}$$

where $W_P \in R^{d\times 1}$ and $b_P \in R^n$ are learnable parameters. In this manner, we also obtain the visual embedding $I^{'} = [i_1^{L^I}, i_2^{L^I}, \cdots, i_r^{L^I}]$ as the $L^I$-th layer outcome in the GAT. Then, the congruity score $s_b$ between $T^{''}$ and $I^{'}$ is computed as follows:

$$G_b = \frac{1}{\sqrt{d}}\left(T^{''} I^{'\top}\right) \tag{9}$$

$$s_b = \text{softmax}\left(T'' W_b + b_b\right)^\top G_b \tag{10}$$

where $G_b \in R^{(n+1)\times r}$ is the similarity matrix between textual and visual modalities, and $W_b \in R^{d\times 1}$ and $b_b \in R^{n+1}$ are learnable parameters.

#### 3.2.2. Global-Level Congruity

As pointed out in the Introduction, there are two main challenges in multimodal sarcasm detection: (a) incomplete visual semantics caused by object detection approaches or patch utilization; (b) the confusion of incongruity identification due to the semantic gap between modalities. Inspired by the work of Xu et al. [5], a global congruity module is devised. In this module, the overall similarity between the text and the image is computed. Then, the feature representations of both textual and visual modalities are optimized by using contrastive learning, with the aim of bridging the semantic gap.

A contrastive learning strategy is proposed to deal with the global-level congruity. Positive and negative examples are generated from an input pair $(X_{cls}, V_{cls})$ within a batch of size $N$. We define $X_{cls}$ and $V_{cls}$ as the sentence representation and image global representation of the Feature Extractor. At this stage, all positive examples are textual and visual representations from the same input pair $(X^a_{cls}, V^b_{cls})_{a=b}$, while negative examples are those from different input pairs $(X^a_{cls}, V^b_{cls})_{a\neq b}$. In this way, each input pair in a batch contains 1 positive example and $N$-1 negative examples. Each example $(X^a_{cls}, V^b_{cls})$ is sent to an MLP layer for transformation into the feature representations $T^a_{cls}$ and $I^a_{cls}$. Both the image-to-text contrastive loss function and the text-to-image contrastive loss function are minimized, based on which the similarity between positive examples is maximized and the similarity between negative examples is minimized [20]. Specifically, the image-to-text contrastive loss function of the $i$-th positive example in a batch can be written as follows:

$$L_i^{(I_{cls}\to T_{cls})} = -\log \frac{\exp\left(\text{sim}\left(I^i_{cls}, T^i_{cls}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(I^i_{cls}, T^j_{cls}\right)/\tau\right)} \tag{11}$$

where $\text{sim}\left(I^i_{cls}, T^i_{cls}\right) = \frac{{I^i_{cls}}^\top T^i_{cls}}{\|I^i_{cls}\|\|T^i_{cls}\|}$ denotes the cosine similarity between $I^i_{cls}$ and $T^i_{cls}$, and $\tau$ is the temperature hyperparameter.

Likewise, the text-to-image contrastive loss function of the $i$-th positive example is

$$L_i^{(T_{cls}\to I_{cls})} = -\log \frac{\exp\left(\text{sim}\left(T^i_{cls}, I^i_{cls}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}\left(T^i_{cls}, I^j_{cls}\right)/\tau\right)} \tag{12}$$

The final loss function for the batch is

$$L_{glo} = \frac{1}{N}\sum_{t=1}^{N}\left(\lambda_a L_i^{(I_{cls}\to T_{cls})} + (1-\lambda_a) L_i^{(T_{cls}\to I_{cls})}\right) \tag{13}$$

where $\lambda_a \in [0, 1]$ is a hyperparameter.

### *3.3. Semantic-Enhanced Module*

As a sentiment-analysis-related task, the result based on only the identification of text–image incongruity may be unreliable. Notably, current sentiment analysis studies reveal that the exploitation of semantics provides a deep-level understanding, which substantially benefits the sentiment polarity prediction [9,10,21]. For this reason, an Intra-modal GNN is designed to capture the semantic information within the sentence. Firstly, an attention matrix $A$, as the adjacency matrix, is derived via the multi-headed self-attention mechanism, which is

$$A = \text{softmax}\left(\frac{(TW_k)(TW_q)^\top}{\sqrt{d_{att}}}\right) \tag{14}$$

$$d_{att} = \frac{d}{K} \tag{15}$$

where $T \in R^{n \times d}$ stands for the textual features generated by the Feature Extractor, $K$ is the head number, and $W_k$,$W_q \in R^{d \times d_{att}}$ are trainable weight matrices. Then, a convolutional neural network (GCN) is applied to extract the semantic information $F$:

$$F = \text{GCN}\left(A, T, W_f\right) \tag{16}$$

where $W_f$ is the trainable parameter matrix of the GCN.

In practical use, people generally convey the same sentiment in a similar manner. As a result, we can also leverage samples with similar expressions to obtain the semantic information. Specifically, supervised contrastive learning can be used to optimize textual feature representation by spatially aggregating semantic-related sentences while separating the unrelated ones [22]. For the construction of positive and negative examples, textual representations with the same label in a batch are positive examples, and those with different labels are negative examples. Let $D = \left(X_{clsi}, X^+_{clsi}\right)$ be an example pair, where $X_{cls}$ is the sentence representation of BERT in the Feature Extractor, and $X_{clsi}$ and $X^+_{clsi}$ are semantic-related texts with the same label. By feeding $X_{clsi}$ and $X^+_{clsi}$ to the MLP, we thus obtain $h_i$ and $h^+_i$ to facilitate the computation. Then, the training objective of the $N$-size batch is

$$L_{sim} = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}\left(h_i, h^+_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(h_i, h^+_i\right)/\tau\right)} \tag{17}$$

where $\tau$ is the temperature hyperparameter, and $\text{sim}(h_1, h_2) = \frac{{h_1}^\top h_2}{\|h_1\|\|h_2\|}$ represents the cosine similarity.

### *3.4. Training and Learning Objectives*

At this point, the token-level congruity (i.e., atomic-level congruity score $s_a$ and composition-level congruity score $s_b$) and the semantic information from the Intra-modal GNN are fused for sarcasm detection. The final sarcastic representation $f$ is defined as follows:

$$p_y = \text{softmax}(IW_i + b_i) \tag{18}$$

$$f = p_y \odot s_a \parallel p_y \odot s_b \parallel F \tag{19}$$

where $W_i \in R^{d1}$ and $b_i \in R^r$ are trainable parameters, $p_y \in R^r$ is the attention vector, and $\odot$ stands for the operation of the element-wise vector product.

The final sarcastic representation $f$ is sent to a fully connected layer with a softmax function. The probability distribution of $f$ in the sarcasm decision space is given as

$$\hat{y} = \text{softmax}(W_o y + b_o) \tag{20}$$

where $W_o \in R^{2 \times 2r}$ and $b_o \in R^2$ are trainable parameters.

Model training is carried out by minimizing the total loss function $L$ with the standard gradient descent algorithm:

$$L_{sar} = -\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \tag{21}$$

$$L = L_{sar} + \alpha L_{sim} + \beta L_{glo} \tag{22}$$

where $\hat{y}_i$ is the prediction outcome of sample $i$, $y_i$ is the real label of sample $i$, $N$ is the size of the training dataset, and $\alpha$ and $\beta$ are hyperparameters to control the contribution of different loss functions.

## 4. Experiments

### *4.1. Dataset*

Experiments were carried out on a publicly available benchmark dataset [13] for multimodal sarcasm detection. This dataset contains tweets of sarcastic expressions as positive examples and those of non-sarcastic expressions as negative examples. Each sample in the dataset consists of text and an attached image. More details of the dataset are shown in Table 1 [2].

**Table 1.** Statistics of dataset.

| | Training | Development | Testing |
|---|---|---|---|
| Positive | 8642 | 959 | 959 |
| Negative | 11,174 | 1451 | 1450 |
| All | 19,816 | 2410 | 2409 |

### *4.2. Experimental Settings*

All data preprocessing procedures were conducted following the method reported by Cai et al. [13]. Every single word in the text and each patch in the image were mapped to 768-dimensional embeddings using pretrained BERT and pretrained ViT, respectively. The layer number of the GAT was set to 2, which is the optimal value according to ongoing studies. The hyperparameters of the loss function, i.e., $\alpha$ and $\beta$, were 0.02 and 0.05. The temperature hyperparameter $\tau$ of simcse was 0.07, while the temperature hyperparameter of text–image global-level congruity computation was 0.17. The Adam optimizer was adopted with a learning rate of 0.0001 and a batch size of 32. To prevent overfitting, both dropout and early stopping were employed. The accuracy and F1-score were taken as metrics to demonstrate the working performance.

### *4.3. Baseline*

The working performance of our model was evaluated in comparison with the following baselines:

(1) Image-based methods: These models exploit only visual information for sarcasm detection, including **Image** [13], which uses **ResNet** [23] for sarcasm classifier training, and **ViT** [18], which applies a pretrained ViT [cls] token representation to detect sarcasm.

(2) Text-based methods: These models exploit only textual information for sarcasm detection, including **TextCNN** [24], which was devised on the basis of a CNN; **Bi-LSTM** [25], which is a bidirectional long short-term memory network for text classification; **SIARN** [26], which uses inner-attention for text sarcasm detection; **SMSD** [27], which captures the textual incongruity by searching a self-matching network; and **BERT** [17], which is a pretrained model with the input "[CLS]text[SEP]".

(3) Multimodal methods: These models leverage both textual and visual information for sarcasm detection, including **HFM** [13], which is a hierarchical multimodal feature fusion model for multimodal sarcasm detection; **D&R Net** [5], which is a decomposition and relation network for both cross-modality contrast and semantic association modeling; **Res-BERT** [6], which is proposed for sarcasm detection by concatenating visual and textual features; **Att-BERT** [6], which explores inter-modal attention and co-attention to model multimodal incongruity; **InCrossMGs** [7], which is a graph-based model and exploits both intra- and inter-modal sarcasm relations; a non-external-knowledge-variant of **CMGCN** [8],

which extracts visual features via object detection and fuses textual and visual information using a cross-modal graph; and **HKE** [2], which is a hierarchical model for sarcasm detection by reasoning atomic-level congruity and composition-level congruity.

## 5. Experimental Results

### *5.1. Main Results*

The experimental results are presented in Table 2. Among all the methods, the proposed model achieves the best and the most consistent working performance. One can observe that the text-based methods are generally more competitive than the image-based methods. Clearly, the textual modality is taken as a better alternative for capturing sarcastic information. In line with this result, the semantic information can be applied to detect the sarcastic expression as an auxiliary. Furthermore, the multimodal methods consistently outperform the unimodal baselines, indicating the effectiveness of exploiting multimodal information in sarcasm detection. Specifically, our model dominates the state-of-the-art methods in all evaluation settings. There is a considerable performance gap between our model and the baselines. A minimum accuracy gap and F1 gap of 1.09% and 1.87% are observed against HKE, which are significant. With the integration of semantic information, our model shows its superiority in identifying the incongruity between texts and images. It is reasonable to obtain more reliable multimodal information and thus better performance, which is observed in this case.

**Table 2.** Main results.

| Model | | Acc (%) | Pre (%) | Rec (%) | F1 (%) |
|---|---|---|---|---|---|
| Image | Image [13] | 64.76 | 54.41 | 70.80 | 61.53 |
| | ViT [18] | 67.83 | 57.93 | 70.07 | 63.43 |
| Text | TextCNN [24] | 80.03 | 74.29 | 76.39 | 75.32 |
| | Bi-LSTM [25] | 81.90 | 76.66 | 78.42 | 77.53 |
| | SIARN [26] | 80.57 | 75.55 | 75.70 | 75.63 |
| | SMSD [27] | 80.90 | 76.46 | 75.18 | 75.82 |
| | BERT [17] | 83.85 | 78.72 | 82.27 | 80.22 |
| Multimodal | HFM [13] | 83.44 | 76.57 | 84.15 | 80.18 |
| | D&R Net [5] | 84.02 | 77.97 | 83.42 | 80.60 |
| | Res-BERT [6] | 84.80 | 77.80 | 84.15 | 80.85 |
| | Att-BERT [6] | 86.05 | 78.63 | 83.31 | 80.90 |
| | InCrossMGs [7] | 86.10 | 81.38 | 84.36 | 82.84 |
| | CMGCN [8] | 86.54 | - | - | 82.73 |
| | HKE [2] | 87.36 | 81.84 | 86.48 | 84.09 |
| | **SEF (Ours)** | **88.45** | **85.35** | **86.58** | **85.96** |

### *5.2. Ablation Study*

An ablation experiment was conducted to demonstrate the contribution of each component in our model. Three variants of our model were used for comparison: (1) removal of the global-level congruity module ($w/o$ global); (2) removal of supervised contrastive learning from the semantic enhancement module ($w/o$ simcse); and (3) removal of semantics from the semantic enhancement module ($w/o$ semantic). For a fair comparison, we used the same parameter settings for each model.

The results of the ablation study are shown in Table 3. The ablation of semantic information from the semantic enhancement module results in the largest performance drop, indicating the significance of semantics for the multimodal sarcasm detection task. In comparison, the use of global congruity also makes a contribution to the sarcasm detection result. The identification of incongruity among multiple modalities is distinctive. In addition, the withdrawal of supervised contrastive learning ($w/o$ simcse) leads to a marginal performance decrease. The exploitation of sentence semantics from the dataset also benefits sarcasm detection.

**Table 3.** Ablation study results.

| Model | Acc (%) | F1 (%) |
|---|---|---|
| **SEF** | **88.45** | **85.96** |
| *w/o* global | 87.91 | 84.97 |
| *w/o* simcse | 88.20 | 85.37 |
| *w/o* semantic | 87.65 | 84.78 |

### *5.3. Effect of GAT Layer Number*

The effect of the GAT layer number on the sarcasm detection result was investigated; see Figure 4a. Both the accuracy and the F1-score vary in line with the GAT layer numbers. The best result is obtained with an optimal GAT layer of 2. Then, the working performance declines with the increasing layer number. A possible explanation is that the over-smoothing of the GAT makes it challenging to distinguish between the nodes of different modalities.

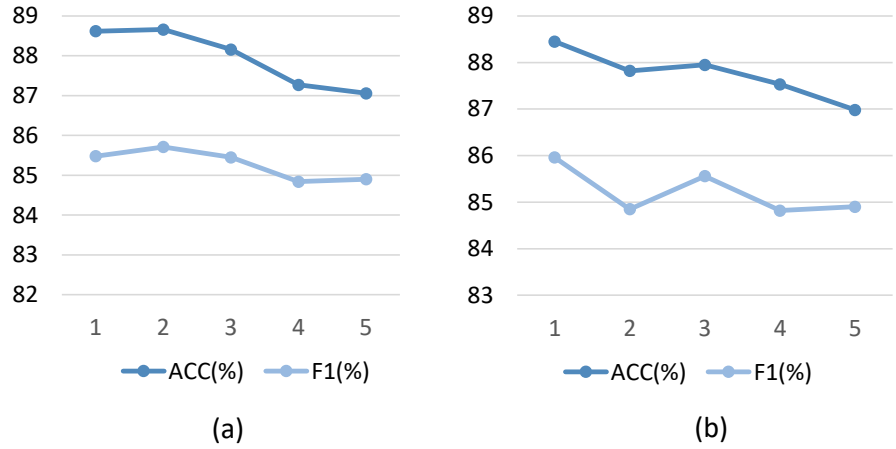


**Figure 4.** (**a**) Results of different GAT layers. (**b**) Results of different GCN layers.

### *5.4. Effect of GCN Layer Number*

Likewise, the effect of the GCN layer number on our model is also reported. As presented in Figure 4b, the best-performing model is available with a GCN layer number of 1. Similar to the GAT, the increase in the layer number also results in a performance drop. With respect to the GCN, the vanishing gradient and information redundancy can arise because of excessive layers, which causes model instability and thus imprecise results.

### *5.5. Case Study*

Three representative cases were selected to validate the effectiveness of the semantic enhancement and global congruity in our model. The sarcasm detection results of **HKE**, **SEF**, and its two variants are presented in Table 4.

In the first two cases, the text intuitively conveys the same sentiment as the image, but their true labels are sarcasm. In the first case, both the text and the image involve drinking a lot of beer, expressing a consistent sentiment according to each modality. Specifically, **HKE** and the **SEF** *w/o* **semantic** identify it as non-sarcasm based on a shallow analysis. By contrast, with the application of semantic information, the **SEF** and **SEF** *w/o* **global** can correctly predict the label as sarcasm. With respect to the second case, the text and the image depict a shortage of potatoes. With the integration of semantic enhancement and global congruity computation, our model predicts the label correctly. These two cases demonstrate the efficacy of semantic information in predicting and facilitating multimodal sarcasm detection.

In the last case, the text delivers a not-big-enough space, while the image illustrates a large space. Its true label is sarcasm due to the inconsistency between the text and the image. However, **HKE** and the **SEF** *w/o* **global** fail to predict the label because the image is divided into multiple patches using ViT. For this reason, the size of the space is invisible, which confuses the computation of text–image incongruity. By contrast, our model is capable of identifying the global congruity of the image, based on which a reliable decision is made.

**Table 4.** Results of three cases using HKE, SEF, SEF *w/o* semantic, and SEF *w/o* global.

| **Methods** | **Effectiveness of Semantic Enhancement** | | **Effectiveness of Lobal Congruity** |
|---|---|---|---|
| |  |  |  |
| | another night having to grind out belgian beer styles , studying for certified <user>. bloody nightmare #beer #nightmare emoji_156 | apparently we have a potato shortage in rotherham this is what i received in a large fries box tonight <user> #valueformoney | hi there <user>, i don't believe this room is large enough for one on one podcasts. #dominion |
| **HKE** | 56 | 56 | 56 |
| **SEF** | 52 | 52 | 52 |
| **SEF** *w/o* **semantic** | 56 | 56 | 52 |
| **SEF** *w/o* **global** | 52 | 56 | 56 |

### *5.6. Visualization*

The textual features extracted by the simcse module are visualized using the t-SNE algorithm [28], aiming at verifying the effectiveness of simcse for semantic enhancement; see Figure 5. For both HKE and the SEF without simcse, the features of the same labels disperse within the space. In comparison, the feature vectors are more aggregated in the SEF. With the application of simcse, not only is the distribution of same-labeled features more concentrated, but the distinguishing-labeled features are also separated to a large extent. As a result, our model shows its capability in textual feature learning and sarcasm detection.

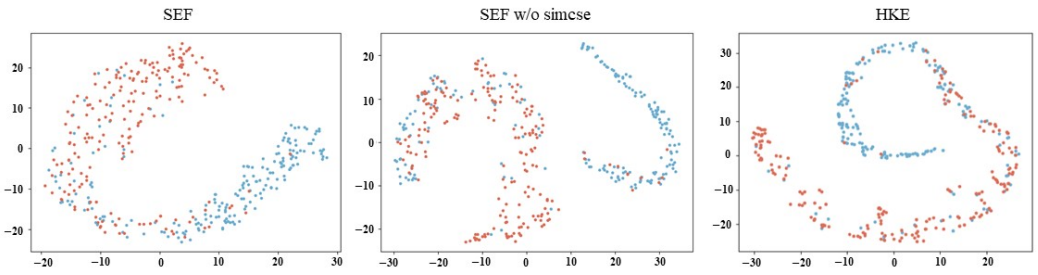


**Figure 5.** Visualization of textual feature vectors. Dots in red and blue represent the non-sarcastic and sarcastic samples, respectively.

## 6. Conclusions

In this work, a semantic enhancement framework is proposed for the task of multi-modal sarcasm detection. To deal with the multimodal relation, the image–text congruity

is modeled on both a multi-scale and multi-span token level. Furthermore, instead of identifying the sarcasm solely based on multimodal incongruity, an Intra-modal GNN is devised to capture the textual semantic information as an auxiliary. The exploitation of textual semantics enhances the semantic delivery via multiple contrastive learning and mitigates cross-modal semantic disparities. Experiments were carried out to verify the superiority of our model and the effectiveness of semantics in multimodal sarcasm detection, which establishes strong evidence of the remarkable working performance. Our model outperforms the state of the art by 1.87% in terms of the F1-score and 1% in terms of accuracy.

However, the results of this paper are subject to certain limitations. Currently, there is a scarcity of datasets for multimodal sarcasm, and the available data are limited in quantity. Expanding the dataset with additional samples would enhance the generalizability of the testing methods.

Additionally, further exploration can be conducted by integrating the proposed approach with large-scale language models for more in-depth analysis to improve the accuracy in detecting nuanced forms of sarcasm across diverse contexts.

**Author Contributions:** Conceptualization, W.Z. and Z.Z.; methodology, W.Z.; formal analysis, W.Z. and Z.Z.; writing—original draft preparation, W.Z. and Z.Z.; writing—review and editing, Y.X. and Q.W.; supervision, Y.X. and Q.C.; funding acquisition, Q.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011370, the National Natural Science Foundation of China (32371114), and the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049).

**Data Availability Statement:** Data are contained within the article.

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

## References

1. Gibbs, R.W. On the psycholinguistics of sarcasm. *J. Exp. Psychol. Gen.* **1986**, *115*, 3. [CrossRef]
2. Liu, H.; Wang, W.; Li, H. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. *arXiv* **2022**, arXiv:2210.03501.
3. Babanejad, N.; Davoudi, H.; An, A.; Papagelis, M. Affective and contextual embedding for sarcasm detection. In Proceedings of the 28th International Conference on Computational Linguistics, Virtual, 8–13 December 2020; pp. 225–243.
4. Kelishadrokhi, M.K.; Ghattaei, M.; Fekri-Ershad, S. Innovative local texture descriptor in joint of human-based color features for content-based image retrieval. *Signal Image Video Process.* **2023**, *17*, 4009–4017.
5. Xu, N.; Zeng, Z.; Mao, W. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 3777–3786.
6. Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; Wang, W. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual, 16–20 November 2020; pp. 1383–1392.
7. Liang, B.; Lou, C.; Li, X.; Gui, L.; Yang, M.; Xu, R. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4707–4715.
8. Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; Xu, R. Multi-modal sarcasm detection via cross-modal graph convolutional network. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1: Long Papers, pp. 1767–1777.
9. Pang, S.; Xue, Y.; Yan, Z.; Huang, W.; Feng, J. Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Virtual, 1–6 August 2021; pp. 2627–2636.
10. Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; Hovy, E. Dual graph convolutional networks for aspect-based sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual, 1–6 August 2021; Volume 1: Long Papers, pp. 6319–6329.
11. Yu, H.; Lu, G.; Cai, Q.; Xue, Y. A KGE Based Knowledge Enhancing Method for Aspect-Level Sentiment Classification. *Mathematics* **2022**, *10*, 3908. [CrossRef]

12. Schifanella, R.; De Juan, P.; Tetreault, J.; Cao, L. Detecting sarcasm in multimodal social platforms. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 September 2016; pp. 1136–1145.
13. Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2506–2515.
14. Yue, T.; Mao, R.; Wang, H.; Hu, Z.; Cambria, E. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. *Inf. Fusion* **2023**, *100*, 101921.
15. Qiao, Y.; Jing, L.; Song, X.; Chen, X.; Zhu, L.; Nie, L. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 9507–9515.
16. Wen, C.; Jia, G.; Yang, J. DIP: Dual Incongruity Perceiving Network for Sarcasm Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2540–2550.
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2010**, arXiv:2010.11929.
19. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
20. Xu, B.; Huang, S.; Sha, C.; Wang, H. MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual, 21–25 February 2022; pp. 1215–1223.
21. Zhu, Z.; Zhang, D.; Li, L.; Li, K.; Qi, J.; Wang, W.; Zhang, G.; Liu, P. Knowledge-guided multi-granularity GCN for ABSA. *Inf. Process. Manag.* **2023**, *60*, 103223. [CrossRef]
22. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Kim, Y. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1746–1751. [CrossRef]
25. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [PubMed]
26. Tay, Y.; Tuan, L.A.; Hui, S.C.; Su, J. Reasoning with sarcasm by reading in-between. *arXiv* **2018**, arXiv:1805.02856.
27. Xiong, T.; Zhang, P.; Zhu, H.; Yang, Y. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2115–2124.
28. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.