Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection

Hao Liu[a], Runguo Wei[a], Geng Tu[a], Jiali Lin[b], Cheng Liu[a], Dazhi Jiang[a,c,*]

[a] Department of Computer Science, Shantou University, Shantou, China
[b] Bussiness School, Shantou University, Shantou, China
[c] Guangdong Provincial Key Laboratory of Information Security Technology, Guangzhou, China

ARTICLE INFO

Keywords:
Multimodal sarcasm detection
Sentiment
Hierarchical fusion network

ABSTRACT

Sarcasm is a form of sentiment expression that highlights the disparity between a person's true intentions and the content they explicitly present. With the exponential increase in multimodal data on social platforms, the detection of sarcasm across various modes has become a pivotal area of research. Although previous studies have extensively examined multimodal feature extraction, fusion, and the modeling of inter-modal incongruities, they often neglected the subtle sentiment cues inherent in sarcastic multimodal data. Additionally, they did not adequately address the sparse distribution and tenuous connections between sarcastic features both within and cross modalities. To address these gaps, we introduce a hierarchical fusion model that integrates sentiment information for enhanced multimodal sarcasm detection. Specifically, we use attribute-object matching in the image modality, treating it as an auxiliary attribute modality. Sentiment data is then extracted from each modality and combined to achieve a more comprehensive representation within modalities. Moreover, we characterize the relationships of inter-modal incongruities using a crossmodal Transformer. We also implement a sentiment-aware image-text contrastive loss mechanism to synchronize the semantics of images and text better. By intensifying these alignments, our model is better equipped to understand incongruous relationships. Experiments demonstrate that our hierarchical fusion model achieves state-of-the-art performance on the multimodal sarcasm detection task.

1. Introduction

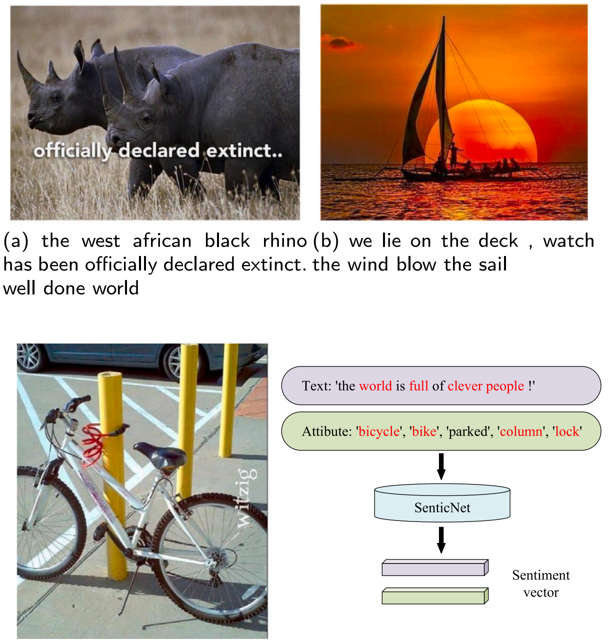
Sarcasm constitutes a distinct form of expression within everyday communication. Beneath the explicit expressions, the true sentiments and intentions of individuals often lie concealed [1,2]. As a result, sarcasm detection plays a crucial role in enhancing the efficacy of sentiment recognition, topic detection, question-answering systems, and opinion mining [3–8]. By deciphering the genuine meanings behind sarcastic comments, these systems can achieve a more profound comprehension of users' actual feelings and perspectives. This, in turn, facilitates more nuanced and insightful results [9].

With the rapid evolution of social networks, a growing number of individuals are avidly sharing daily experiences and participating in discussions on current events across platforms like Facebook, Twitter, and other major social networks. This active engagement has led to the generation of vast amounts of multimodal data. When compared to solely using unimodal data, such as text or image [10–12], multimodal data can furnish more holistic and precise insights [13–17]. Indeed, amalgamating text and image information can offer a more accurate reflection of an individual's genuine feelings and intentions. Consider Fig. 1(a), where the text *'the west African black rhino has been officially declared extinct. Well done world'* conveys disdain for the extinction of the black rhino. This sentiment is incongruous with the accompanying image. In contrast, Fig. 1(b) contains a non-sarcastic expression: *'we lie on the deck, watch the wind blow the sail'*, which is congruent with the depicted scene. Such examples highlight the need to discern any misalignment between text and image content to determine if sarcasm is present [18]. A significant challenge in multimodal sarcasm detection is effectively extracting and merging information from different modalities while understanding the inter-modal incongruities.

For multimodal sarcasm detection from image and text, some previous works fuse modal features by concatenating extracted text and image features during early fusion [13]. Alternatively, attention mechanisms are employed to fuse multimodal features later [19,20]. Some researchers extract crossmodal information through attention [21]. Recently, interactive graph networks were utilized to model the relationship between modalities and facilitate learning intra-modal incongruity [22]. Although these methods demonstrate robust sarcasm

[*] Corresponding author at: Department of Computer Science, Shantou University, Shantou, China.
E-mail address: dzjiang@stu.edu.cn (D. Jiang).

https://doi.org/10.1016/j.inffus.2024.102353
Received 6 November 2023; Received in revised form 10 January 2024; Accepted 11 March 2024
Available online 13 March 2024


officially declared extinct..
witzig
Text: 'the world is full of clever people !'
Attibute: 'bicycle', 'bike', 'parked', 'column', 'lock'
SenticNet
Sentiment vector

(a) the west african black rhino has been officially declared extinct. well done world

(b) we lie on the deck , watch the wind blow the sail

(c) The words marked in red are sentimental words. The attribute words denote the local information of the image.

Fig. 1. (a) and (b) are two multimodal examples of the Twitter sarcasm dataset. (a) Sarcasm example; (b) Non-sarcasm example. (c) the illustration for the text, image, attribution modalities, and sentiment vector.

detection, there are some limitations: (1) Sarcasm detection is sentiment-related, but current research often fails to incorporate sentimental information during feature extraction adequately. (2) Image and text semantics are not fully aligned. Visual information related to sarcasm tends to be sparsely distributed [15], resulting in a weak correlation between modalities.

Given the aforementioned challenges, simply merging multimodal features does not effectively promote information exchange between different modalities to identify incongruities between them. To address this, the conceptual content of an image can be harnessed as an attribute modality. This, coupled with suitable modal interaction techniques, can amplify the connection between text and image content. As depicted in Fig. 1(c), the attribute serves as an abstraction of the image, focusing on its vital local details. The words highlighted in red in both the text and attribute modalities are sentiment-laden, enhancing the representation within a singular modality. As a result, multimodal sarcasm detection requires the formulation of intricate interrelationships between text and imagery.

This paper introduces the Multimodal Sentiment-Aware Hierarchical Fusion Network (SAHFN) designed to tackle sarcasm detection tasks. It achieves this by integrating sentimental data and modeling interdependencies among various modalities. Initially, text and attributes are segmented into individual words to further infuse sentimental information. Word-level sentiment scores are derived from SenticNet7 [23], forming an sentimental vector from these scores. Subsequently, by employing a sentiment-aware attention mechanism, intra-modal and sentiment features are merged to achieve a representation for each modality. Both sentiment-infused and raw features contribute to the computation of the Sentiment-Aware Text-Image Contrastive Loss. Aligning text and image semantics through this Contrast Loss bolsters inter-modal correlation, enhancing the model's proficiency in discerning both sentiment and semantic information across modalities. Such an enhancement primes the model for a deeper understanding of inter-modal incongruities. To capture dependencies among modalities and better understand these incongruities, we integrate three crossmodal transformers. Additionally, to strengthen the connection between text and images, we employ five tokens from each Twitter image as the attribute modality. The final prediction vector is then refined using an attention mechanism. Experiments conducted on a publicly available dataset demonstrate that our proposed model effectively leverages multimodal information, significantly improving sarcasm detection performance.

We sum up the contributions of this article as below.

- We introduce sentiment information into the model to obtain sentiment-aware intra-modal representations.
- For the first time, we introduce a modal contrastive loss during sarcasm detection model training. This aligns image and text semantics to enhance correlation and the model's ability to extract inter-modal incongruous information. Crossmodal transformers are also used to model inter-modal dependence and learn incongruity.
- We conduct extensive experiments and analysis, demonstrating the state-of-the-art performance of our model on a public dataset.

2. Related work

2.1. Sarcasm detection with text

Sarcasm recognition is a type of emotion analysis research [24–26]. Traditional sarcasm detection research has focused on text modality data [27,28]. Text-based sarcasm detection can be categorized into context-free and context-related [29]. Context-free detection analyzes only the target sentence [30] without context. Context-related detection analyzes the target and its context [31–33]. Tay et al. [34] presented an attention-based network that enables modeling common sarcasm incongruities, outperforming baselines on Twitter [35] and Reddit [36] datasets. Besides, some researchers focus on and utilize transfer learning to transfer emotion-related resources to irony detection [37]. Sarcasm judgment relies heavily on broad context like speaker state, text position, social media comments, etc. [38]. Poria et al. [39] used pre-trained models to extract personality and emotion features for sarcasm recognition. Hazarika et al. [40] combined multiple contexts in CASCADE to identify sarcasm using complex contextual information. They considered posts as content and context (user and subject information). Alex Kolchinsk [41] focused on user information by simplifying context, assuming different user expression habits. Compared to the complex context in CASCADE, this achieved better performance. BERT has also been used to embed sentiment and context for textual sarcasm detection [42]. However, single-modal data no longer meets sarcasm detection needs.

2.2. Multimodal sarcasm detection

The primary objective of the multimodal sarcasm detection task is to predict ironic labels by harnessing and amalgamating information from multiple modalities. Analyzing sarcasm cues within multimodal data, which includes sound, images, text, and video, and effectively utilizing emotional information, poses a significant challenge in the current field of sarcasm detection [43–45]. In comparison to unimodal scenarios, multimodal scenarios present a more intricate information landscape where modalities complement each other. Rosso et al. [46] discuss the role of different modalities in multimodal irony detection. Early research tackled multimodal sarcasm identification tasks by manually constructing features. For instance, Schifanella [13] employed deep neural networks to extract visual and text features, which were then fused using SVM to classify sarcasm labels. Cai et al. [19] curated a Twitter multimedia sarcasm dataset and introduced a straightforward hierarchical fusion network. Subsequently, Wang [47] and Ashraf [48] utilized BERT [49] as the text encoder and extracted intra-modal information through attention networks. Rosso et al. [50] utilizes the

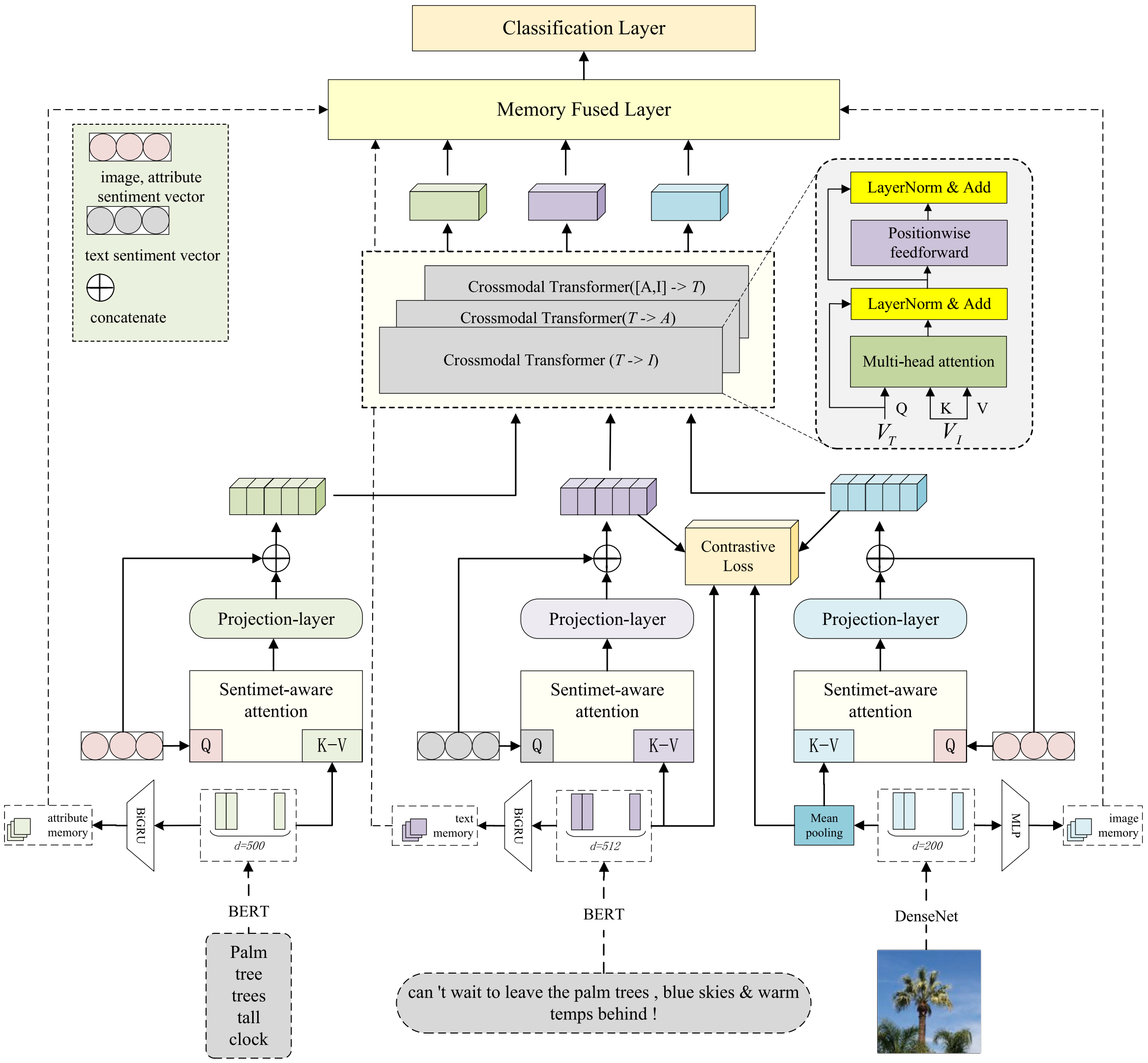


Fig. 2. The overall framework of our proposed SAHFN.

transformer-based architecture to extract text and image features and concatenate them as a fusion representation for irony recognition. Liang et al. [22] obtained single-modal features by constructing individual single-modal graphs and constructed crossmodal interaction graphs for multimodal interaction. Recognizing the complexity of image data, Liang [51] opted to use extracted visual area information instead of the entire image's features. Furthermore, Castro et al. [43] proposed multimodal sarcasm recognition in dialogue tasks, fusing multimodal data by concatenating text, audio, and video features. Pramanick [52] incorporated multimodal information by employing inter-modal and intra-modal attention mechanisms. Liu [53,54] employ multi-view and common knowledge to identify sarcasm labels. Zhang et al. [55] leveraged fuzzy networks to tackle the task. In this paper, our focus centers on image-text sarcasm detection.

3. Methodology

3.1. The overall framework

The proposed multimodal sarcasm detection framework (SAHFN) consists of three main components: Encoding Layer, Fusion Layer, and Memory Fusion Layer. The overall framework is exhibited in Fig. 2. The encoding layer is employed to extract image, text, and attribute features. The Fusion Layer is divided into two distinct components: the sentiment-aware layer and the multimodal interaction layer. Each sample is divided into three modalities: image, text, and attribute. The raw feature vectors from these three modalities are extracted within the Encoding Layer. Subsequently, the raw feature vectors are fed into the next layer for fusing sentiment-related information and facilitating multimodal interaction. The representation of each modal obtained through the multimodal interaction layer is ultimately fused with the memory vector to obtain the vector for sarcasm detection. The details are as follows.

3.2. Feature extraction layer

3.2.1. Image feature representation

In our work, the pre-trained DenseNet-121 [56] model is utilized to obtain the representation of Tweet images. The input tweet image is resized to 448 × 448 and evenly divided into 14 × 14 regions and gets the raw image feature vector of each region $(i = 1, 2, 3, \dots, 196)$ through DenseNet-121:

$$r_{region_i} = DenseNet121\left(I_i\right) \tag{1}$$

Subsequently, we conduct an average pooling operation on the image feature vector of all regions of the image to get the raw image feature vector:

$$r_{image} = \frac{1}{N}\sum_{i=1}^{N} r_{region_i} \tag{2}$$

where N is the number of areas divided by the image, which is set to 196. The guidance image feature is stacked by region image feature. $g_{image} = \left\{ r_{region_i} \right\}^n$, $n = 196$.

3.2.2. Text and attribute feature representation

Attributes can be extra knowledge to enhance the connection between images and text. The extracted attributes can describe the background and potential information of the image. These attributes are treated as an extra modality. We employ the pre-trained ResNet model as the image attribute prediction model. The ResNet is pre-trained on COCO datasets with 1000 classes. Then, we feed each image into ResNet to predict the multi-labels. Among that, we use five labels with the highest confidence score as the image's attribute modality as image's attribute modality. We extract their representation as the attribute modal vector at the word level.

$$Attribute = ResNet\left(I\right) \tag{3}$$

where I is a image. $Attribute = \left\{w_i\right\}_n$, $n = 5$. w_i is the attribute label of image.

Previous studies have shown that most of the sarcasm expressions in the text appear in an asymmetric context in the text [57]. Therefore, we leverage the pre-trained model BERT [49] to extract text feature. We can obtain the word-level context vector representation of the sentences by superimposing the Token Embedding, Segment Embedding, and Positional Embedding of the word. Besides, using the superimposed embeddings as the sentence input effectively solves the complex problem of polysemy of a word in the text. From the outputs of the BERT model, we utilize both the last_hidden_state and pooler_output. The last_hidden_state is regarded as the raw feature x_i and the pooler_output is regarded as guidance feature g_m, $m \in \{text, attribute\}$. Besides BERT, the pre-trained model RoBERTa [58] is also employed to extract text and attribute modal features.

After extracting three modal features, for each modality m, $m \in \{text, image, attribute\}$, raw features r_m are fed into fused layer. The guidance features g_m are fed into three simple networks to gain the memory vectors represented as R_m.

3.3. Fusion layer

The Fusion Layer consists of two sublayers: the Sentiment-Aware Layer and the Multimodal Interaction Layer.

3.3.1. Sentiment-aware layer

For text modality data, the Natural Language Toolkit (NLTK) [59] is leveraged to split the sentence S into words $\left[w_1, w_2, \dots, w_n\right]$, For each word w_i, the sentiment score c_i is obtained by querying the sentiment lexicon SenticNet7 [23] at word-level. All the sentiment scores form a sentiment vector as the text sentiment embedding C_{text}. Similarly, for the attribute modality, we gain the attribute sentiment embedding $C_{attribute}$ using the identical approach as that employed for the text modality.

For each modality m, $m \in \{text, image, attribute\}$, sentiment embedding C_m and raw feature r_m are fused through sentiment-aware attention. We map the sentiment vector to Query (Q), and the raw feature r_m is mapped into Key (K) and Value (V). Through the sentiment-aware attention, the model could capture the intra-modal semantic information while fusing the sentimental information to derive the sentiment-aware features, which are denoted as X:

$$Q_m^i = W_m^Q \cdot C_m^i + b_m^Q \tag{4}$$

$$K_m^i = W_m^K \cdot r_m^i + b_m^K \tag{5}$$

$$V_m^i = W_m^V \cdot r_m^i + b_m^V \tag{6}$$

$$y_m^i = softmax\left(\frac{K_m^{iT} \cdot Q_m^i}{\sqrt{d}}\right) \cdot V_m^i \tag{7}$$

$$X_m^i = W_m \cdot C_m^i + projection\left(y_m^i\right) \tag{8}$$

where W_m is trainable parameter. The $projection$ is a mapping layer $(Linear{-}Relu{-}Linear)$. $softmax\left(\cdot\right)$ calculates the sentiment-aware score matrix of the raw feature r_m, representing the attention of the raw feature r_m to sentiment.

3.3.2. Multimodal interaction layer

We employ the crossmodal transformer to interact with the information of each modality to learn the dependencies relationship between modalities. In this layer, three modalities interact with the three transformers. Within the crossmodal transformer, crossmodal attention is employed to fuse different modal features. The vector X_α and modality X_β, which denote two different modal features obtained by the sentiment-aware layer, are fed into the $transformer_\alpha$, where $(\alpha, \beta) \in \{(image, text), (attribute, text), (text, im \oplus at)\}$. The $im \oplus at$ denotes the image modal feature concatenated with the attribute modal feature. Inspired by the work [60,61], we believe that adaption crossmodality effectively fuses multimodal information. So for the input of modality α and modality β, which is fed into Cross Attention(CA) firstly, the queries, keys and values are denoted as $Q_\alpha = W_\alpha^Q \cdot X_\alpha^Q$, $K_\beta = W_\beta^K \cdot X_\beta^K$, $V_\beta = W_\beta^V \cdot X_\beta^V$ respectively. The W_α^Q, W_β^K, and W_β^V are trainable weights. The representation denotes the output of the crossmodal attention:

$$y_\alpha = CA_{\alpha,\beta}\left(X_\alpha, X_\beta\right) = softmax\left(\frac{K_\beta^T \cdot Q_\alpha}{\sqrt{d}}\right) \cdot V_\beta \tag{9}$$

where y_α is the output of each layer in the crossmodal attention. The crossmodal attention has n layers, and the output from each layer is concatenated to obtain the final output through a linear layer:

$$Y_\alpha = W_\alpha \cdot concatenate\left(y_\alpha^1, y_\alpha^2, \dots, y_\alpha^n\right) \tag{10}$$

Then, a position-wise feed-forward layer (ffd) is followed to compute the output. Concurrently, we connect the output with X_α through a residual connection. The last two layers of crossmodal $transformer_\alpha$ are represented by function $f\left(\cdot\right)$. The output of crossmodal $transformer_\alpha$ is computed as follows.

$$o_\alpha = f_\alpha\left(ffd\left(X_\alpha + norm\left(Y_\alpha\right)\right)\right) \tag{11}$$

The multimodal interaction layer has n layers. The output of the last layer is the input of the next layer: $X_m^{l+1} = o_m^l$.

3.4. Memory fused layer

We first extract raw features through three simple networks for each modality and temporarily store them as memory features. The memory features R_m is the intra-modal low-level feature devoid of sentiment information. The g_{text} and $g_{attribute}$ are extracted by BiGRU respectively, and g_{image} is extracted by MLP:

$$R_{image} = MLP\left(g_{image}\right) \tag{12}$$

$$R_\gamma = BiGRU_\gamma\left(g_\gamma\right) \tag{13}$$

where $\gamma \in \{text, attibute\}$.

Next, we calculate the similarity between different modal features using the attention mechanism, dynamically adjust the weight between different modal features, and further capture the correlation between various modal features. Since the raw image feature does not undergo the mean pooling operation, the memory feature of each mode has one more dimension than the feature through the Fused Layer. R_m^i denotes the ith memory vector of m modality, and v_n denotes the vector of n modality through the Fused Layer. The vector S_{mn}^i acquired by concatenating R_m^i and v_n is entered into a feed-forward layer with two sublayers to obtain α_{mn}^i. We calculate the average value of α_{mn}^i as the reconstruction weight of m modality to get the feature vector of each modality:

$$S_{mn}^i = concatenate\left(R_m^i, v_n\right) \tag{14}$$

Table 1
Statistics of the multimodal sarcasm dataset.

	No Sarcasm	Sarcasm	Total	Sentence length
Training	8642	11 174	19 816	16.91
Development	959	1451	2410	16.92
Test	959	1450	2409	17.13

$$\alpha^i_{mn} = W_{mn2} \cdot \ tanh\left(W_{mn1} \cdot X^i_{mn} + b_{mn1}\right) + b_{mn2} \tag{15}$$

$$\alpha^i_m = \frac{\sum_{n \in \varphi} \alpha^i_{mn}}{3} \tag{16}$$

$$v_m = \sum_{i=1}^{Lm} \alpha^i_m \cdot R^i_m \tag{17}$$

where α^i_m denotes the reconstruction weight of m modality, $\varphi = \{text, image, attribute\}$. W_{mn1} and W_{mn2} are trainable weights, b_{mn1} and b_{mn2} are trainable bias. L_m is the length of memory features.

To make the final fusion, firstly, we transform the feature vector v_m into v^o_m with a fixed length, length=512. Then, the v_m is fed into a feed-forward layer with two sublayers to get the attention weights β_m and calculate the vectors v_f:

$$v^o_m = \ tanh\left(W_{m1} \cdot v_m + b_{m1}\right) \tag{18}$$

$$\beta_m = softmax\left(W_{m3} \cdot \ tanh\left(W_{m2} \cdot v_m + b_{m2}\right) + b_{m3}\right) \tag{19}$$

$$v_f = \sum_{m \in \varphi} \beta_m \cdot v^o_m \tag{20}$$

where W_{m1}, W_{m2}, W_{m3}, b_{m1}, b_{m2}, and b_{m3} are all trainable parameters.

3.5. Text-image contrastive loss

The text and image modality vectors obtained through the Sentiment-Aware Layer, along with the guidance feature vector, are used to calculate the Text-Image Contrast Loss ($Loss_{tic}$). Through optimizing $Loss_{tic}$, the Sentiment Aware attention can obtain more mutual information between modalities and align image and text semantic information. The $Loss_{tic}$ is calculated as follows.

$$\mathcal{L}_1 = \log \frac{e^{\mathrm{sim}\left(X^i_T, X^i_I\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(X^i_T, X^j_I\right)/\tau}} \tag{21}$$

$$\mathcal{L}_2 = \log \frac{e^{\mathrm{sim}\left(r^i_T, r^i_I\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(r^i_T, r^j_I\right)/\tau}} \tag{22}$$

$$\mathcal{L}_{tic} = \mathcal{L}_1 + \mathcal{L}_2 \tag{23}$$

where the temperature parameter τ is set to 0.1.

3.6. Sarcasm classification

Finally, a sarcasm classification layer is employed to predict sarcasm labels using the above fusion vector:

$$\hat{y}_i = sigmoid\left(W_o \cdot v_f + b_o\right) \tag{24}$$

After that, we calculate the binary cross entropy loss according to the ground-true labels and predict values, and the model is trained by the joint loss:

$$\mathcal{L}_{sa} = -\frac{1}{n} \sum_{i}^{n} \left[y_i \cdot \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - \hat{y}_i\right)\right] \tag{25}$$

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{sa} + \lambda \cdot \mathcal{L}_{tic} \tag{26}$$

where the n is the number of samples in train sets and λ is Hyper-parameter.

Table 2
Parameter statistics.

Hyper-parameters	Value
Batch size	64
Attribute embedding size	200
Densenet121 FC size	1024
Modality fusion size	512
BERT embedding dimension	768
dropout rate	0.25
Learning rate	0.0002
Weight decay	1e−65
Numbers of fused layer	2
Heads of crossmodal attention	4
Epoch	10
The rate λ of loss	0.38

4. Experiment settings

4.1. Datasets and evaluation metrics

The model presented in this paper is evaluated on the publicly available Twitter multimodal sarcastic dataset constructed by Cai et al. [19]. The dataset is composed of user comments on the Twitter website, including text and relevant images. The dataset is divided into training, development, and test sets. The training set contains 19816 multimodal samples. The development and test sets contain 2410 and 2409 multimodal samples, respectively. Among them, the label 0 means non-sarcasm, while 1 means sarcasm. Table 1 reports the statistical results of the dataset.

Our model is evaluated on another multimodal dataset proposed in the 2022 year, created by Maity [62], called 'multibully.' The dataset consists of two modalities: image and text. The authors collect 5865 samples from social platforms, Twitter and Reddit, with a total of 5 labels: Bully, Sentiment, Emotion, Sarcasm, and Harmfulness. For our task, we focused exclusively on the Sarcasm label. However, some damaged samples are identified in the dataset as published by the authors. After their removal, there are still 5808 samples left in the paper.

Following the previous works, we use the Accuracy score (Acc), Precision (Pre), Recall (Rec), and F1 score (F1) as the evaluation metrics. Meanwhile, due to the uneven distribution of Twitter data sets and to comprehensively evaluate the experimental results, the Macro-averaged Precision (Macro Pre), Recall (Macro Rec), and F1 score (Macro F1) are used as evaluation metrics, too.

4.2. Hyper-parameters settings

This section states some crucial parameters. The pre-trained models BERT-base-uncased [49] and RoBERTa [58] are used to extract text and attribute features. The pre-trained models DenseNet121 [56] are used to extract image features. Meanwhile, we utilize the Adam optimizer to calculate and update the model's parameters. The remaining Hyper-parameters are listed in the Table 2.

4.3. Comparison baselines

Our proposed model is compared with several classical sarcasm detection baselines. The specific description is as follows.

TextCNN [64]: Using convolutional neural network for text irony classification.

TextCNN-LSTM [64]: On the basis of TextCNN, it further extracts text features using LSTM.

SIARN [34]: The inter-word impact is obtained by inter-attention, and then the text sarcasm is detected by short - and long-term memory networks.

BERT [49]: The BERT-base-uncased model is utilized to extract text features as the input of sarcasm tasks.

Table 3
Main experimental results of unimodality and multimodality on Twitter dataset [19]. The result of the baselines can be searched from [22] and [63]. The experimental results getting promoted are in bold.

Modality	Model	Acc (%)	F1-score			Macro-average		
			Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)
$Text$	TextCNN	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	SIARN	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
	SAHFN-BERT (ours)	83.85	79.36	82.32	80.81	83.29	83.62	83.43
	SAHFN-RoBERTa (ours)	92.73	88.81	94.97	91.79	92.42	93.02	92.63
$Image$	Image	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40
	ImageGraph	73.89	63.24	82.17	71.47	–	–	–
$Text + Image$	HFN	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&R Net	84.02	77.97	83.42	80.60	–	–	–
	Attr-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs	86.11	81.38	84.36	82.84	85.39	85.80	85.60
	HCKE	87.02	82.97	84.90	83.92	–	–	–
	SAHFN-BERT (ours)	**87.22**	82.71	**87.33**	**84.95**	**86.71**	**87.23**	**86.92**
	Bridge-RoBERTa	88.51	82.95	89.39	86.05	–	–	–
	FiLMing-RoBERTa	93.66	90.56	93.87	92.19	–	–	–
	SAHFN-RoBERTa (ours)	**96.17**	**95.74**	**95.84**	**95.79**	**96.11**	**98.12**	**96.12**

Table 4
The compare results on multibully dataset [62].

Modality	Model	ACC	F1	Macro F1
$Text$	TextCNN	53.94	47.69	53.27
	TextCNN-LSTM	55.93	50.76	56.00
$Image$	RestNet	55.66	54.21	54.57
$Text + Image$	HFN	55.43	59.71	54.92
	Maity-BERT	58.70	60.12	58.64
	SAHFN-BERT	**59.87**	**63.30**	**59.52**
	Maity-RoBERTa	60.05	64.12	59.54
	SAHFN-RoBERTa	**61.59**	**67.73**	**60.15**

Image [65]: The vector of the image obtained from the pre-trained model ResNet is regarded as the input to predict sarcasm labels through the classification layer.

ViT [66]: The vision pre-trained model ViT extracts the image representation using $[CLS]$ token.

ImageGraph [22]: The static image is constructed for each image. Then, several layers of GCN are used to extract the image's abstract features to capture the image's visual information, which can be employed to predict sarcasm labels.

HFN [19]: Using GloVe and ResNet as feature extractors. Then, adopting a hierarchical fusion network with a single attention layer to fuse the multimodal features.

D&R Net [67]: The text is fused with ANP words through attention, and image and text features are fused using the Decomposition and Relationship Network.

Attr-BERT [68]: A modal is proposed using co-attention to capture the inter-modal incongruity for multimodal sarcasm detection.

InCrossMGs [22]: Firstly, the text and image modal graphs are constructed. Another crossmodal interaction graph is utilized to fuse the two modalities.

HCKE [63]: Leveraging graph neural network to capture the relationship between modalities and analyzing the semantics in sentences for multimodal sarcasm detection.

Bridge-RoBERTa [47]: The embedding of text and image is input into the BERT coding layer together to obtain multimodal joint representation. Finally, features are further extracted for sarcasm detection through 2D attention.

FilMing-RoBERTa [69]: The text features extracted by GRU are fused with image features in different stages to model the inter-modal incongruity.

Maity [62]: It is a baseline evaluated on the multibully dataset using BERT-GRU and ResNet to extract features.

5. Experimental result and analysis

5.1. Main experiment result

The results of our presented model (SAHFN) and other comparison baselines on the Twitter dataset are shown in Table 3, where multimodal data is the input of our experiment. As seen from Table 3, the results obtained by our model perform better in multiple metrics than comparable baselines. When BERT is used as the text feature extractor, the Acc is 87.22%, Rec is 87.33%, and the F1 score is 84.95% of our result. Compared with HCKE, the Acc improved by 0.2%, Rec improved by 2.43%, and F1 improved by 1.03%, respectively. When RoBERTa is used as the text feature extractor, the Acc and f1 are 96.17% and 95.79%, respectively. Compared with FilMing, the Acc improved by 2.51%, and F1 improved by 3.6%, respectively. Additionally, the results of our model on the Macro-average metrics are also better than those of the benchmark, which reveals that SAHFN not only performs well in extracting different modal information but also improves the performance of sarcasm detection through crossmodal information interaction. In addition, we also conduct experiments on 'multibully' dataset. As displayed in Table 4, ACC and F1 are 59.87% and 63.30% when employing BERT as a text extractor, respectively. While employing RoBERTa as a text extractor, the ACC and F1 are 61.59% and 67.73%, respectively. The results indicate that our method performs better than baselines. In the unimodal model, taking the Twitter dataset as an example, the results using text modality are better than those using image modality, which indicates that text can provide more information in sarcasm detection. However, it is notable that the models exhibit outstanding performance when multimodal data is employed as input. As demonstrated in Table 3, our model outperforms all comparable baselines, which shows that embedding sentiment information and using transformer learning the inter-modal incongruity is effective in sarcasm detection tasks. Simultaneously, to explore why the method proposed in this paper can improve the performance of sarcasm detection without paying attention to the feature extractor, the subsequent experimental results come from the experiments using BERT as the text encoder. We also make significance tests with the multimodal model we compared through the t-test and p-value < 0.05.

5.2. Multimodal result analysis

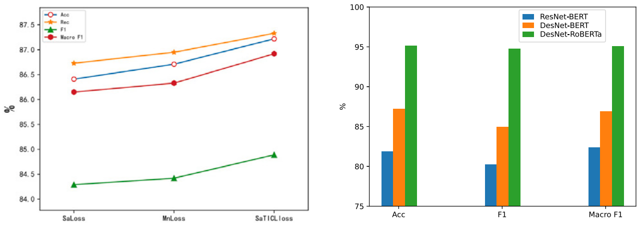
In contrast to unimodal data, such as text or images, multimodal data offers richer information for sarcasm detection and more accurately reflects users' real ideas. To verify the significance of multimodal

Table 5
Experimental results of different modalities with BERT or RoBERTa are used as text extractors on the Twitter dataset. (Legend: B denotes BERT, R denotes RoBERTa.)

Modality	Acc_B	$F1_B$	$MacroF1_B$	Acc_R	$F1_R$	$MacroF1_R$
I	68.12	63.25	69.48	–	–	–
A	66.88	62.25	66.37	87.41	85.50	87.19
T	83.85	80.81	83.43	92.73	91.79	92.63
I+A	72.07	68.86	71.77	91.68	90.43	91.53
T+I	85.75	83.28	85.44	94.35	93.27	94.20
T+A	85.67	83.18	85.35	94.48	93.44	94.33
T+I+A	**87.22**	**84.95**	**86.92**	**96.17**	**95.79**	**96.12**

Table 6
The results of ablation experiments. (Legend: S: sentiment-aware layer, F: multimodal fused layer, M: memory fused layer, L: Sentiment-Aware Text-Image Contrastive Loss).

Model	Acc	F1	Macro F1
SAHFN	**87.22**	**84.95**	**86.92**
w/o L	86.44	84.22	86.17
w/o S	85.43	82.58	85.03
w/o F	86.48	82.79	85.83
w/o M	86.81	84.22	86.44
w/o S-F	84.62	81.86	84.25
w/o S-M	83.85	80.81	83.43
w/o S-F-M	81.41	79.03	83.55



(a)

(b)

Fig. 3. (a) The results of loss analysis; (b) The results of generalizable analysis.

information, we conducted some comparative experiments and analysis on unimodal and multimodal data, and the results are shown in Table 5. To study the impact of image attributes on images and text, we use image attributes as attribute modalities (I). Different modalities interact with information through the crossmodal transformer. The results presented in Table 5 show that the image modality and attribution modality alone have the poorest performance and significantly have a vast gap with the results using text modality, which demonstrates that extracting information related to sarcasm from images is difficult. Besides, results obtained through unimodal data are consistently inferior to those garnered from multimodal data. Moreover, in the case of the T+I and T+A modalities, the results would be basically equal, indicating that the task and scene of describing images using some words can adequately represent the visual information of images. When employing the three modalities (T+I+A) as inputs, the model can fully use the information related to sarcasm in different modalities. Besides, when we adopt Roberta as the text extractor, we can achieve better results in unimodal and multimodal modalities.

5.3. Ablation study

To evaluate the significance of different modules, we present the results of ablation experiments and provide a detailed analysis of the roles of different modules. The results are obtained on the Twitter dataset. In this section, we employ BERT as the text extractor. The modules involved in the ablation experiment are the sentiment-aware layer, multimodal fused layer, and memory fused layer. Table 6 reports the results of the ablation experiment.

Impact of sentiment-aware layer: We remove the sentiment-aware layer component (w/o S) to study its effect on the model. When removing the sentiment-aware layer, the sentiment vector will not be introduced into the model. The sentiment-aware attention is replaced with self-attention. The results in Table 6 show a decline in the model's performance upon removing this layer, which indicates that the model cannot extract more abundant representations of semantic information without sentiment vector embedding.

Impact of multimodal interaction layer: As we can see (w/o F), the result decreases when the multimodal interaction layer is removed. i.e., we remove the crossmodal transformer. The features extracted by the sentiment-aware attention are directly fed into the memory fusion layer for later fusion. We attribute this to the lack of an effective inter-modal interaction mechanism, which makes the model unable to sufficiently learn and capture the incongruity between modalities.

Impact of memory fused layer: To investigate the role of this layer in the model (w/o M), we experiment by directly concatenating the features obtained from the multimodal interaction layer as the predicted features while removing the memory fusion layer. The results exhibit a decline in model performance without memory fusion, which indicates that the memory fused layer enables the model to leverage the low-level information unique to each modality stored in the memory.

Impact of Text-Image Contrastive Loss: We remove the sentiment-aware Text-Image Contrastive Loss (w/o L) and only utilize the cross loss calculated by the predicted labels and the true labels, as shown in formula (25), to train the model. There is a slight decrease in the result. This contrastive loss can align semantic information between images and text to enhance their inter-modal weak correlation.

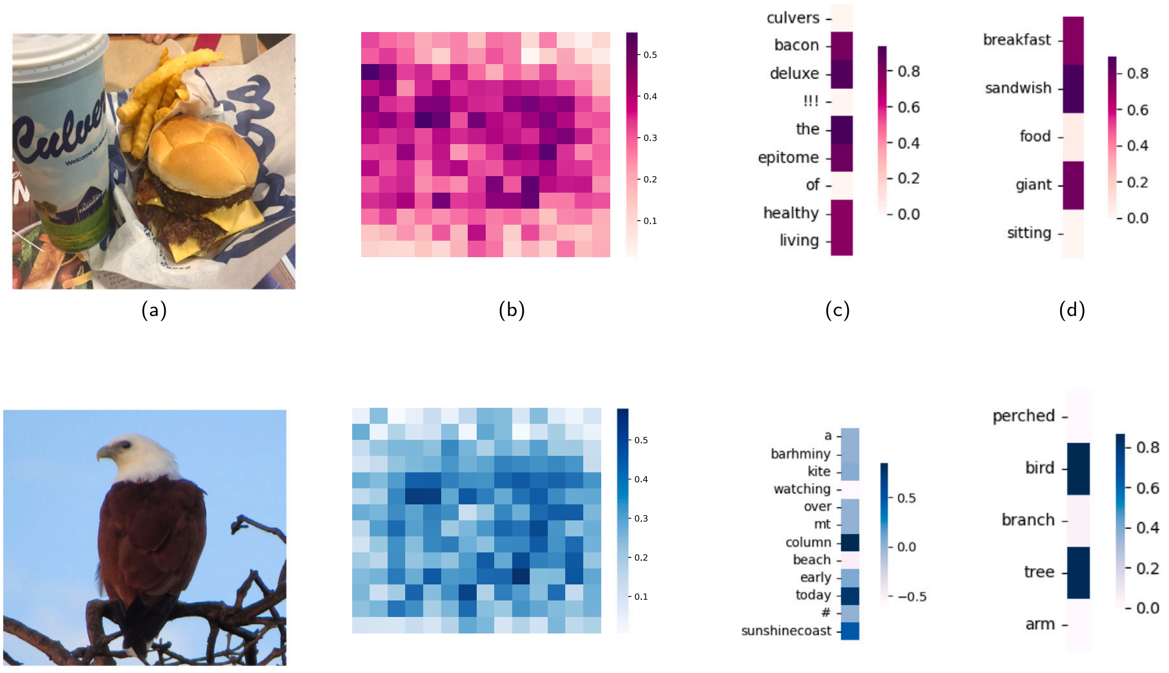
The model's performance decreases when one or more modules are removed, meaning different modules can work collaboratively. In the experiment of removing all layers, the raw feature vectors are concatenated and employed as the input of the classification layer.

5.4. Loss analysis

To study the influence of Text Image Contrast Loss on the proposed model, we conducted three comparative experiments, and the results are illustrated in Fig. 3(a). Each experiment uses a different loss training model. The experimental results are acquired on the Twitter dataset utilizing BERT to extract text features. (1) SaLoss: This loss is computed by the value obtained from the classification layer and the sarcasm labels. (2) MmLoss: This loss is computed for each modality, and the final loss is the sum of the three losses and the SaLoss of the three modalities. (3) STICLLoss: This is a joint loss according to formulate (26). Fig. 3(a) shows a slight improvement in results when utilizing the multimodal loss (MmLoss) in the joint training model. Furthermore, the results are further improved when using the STICLLoss training model as opposed to MmLoss, which shows that Text-Image Contrastive Loss can make the model take into account the training state of different modalities while training to align image and text semantic information to enhance the inter-modal weak correlation. It strengthens the model's ability to extract intra-modal incongruity.

5.5. Generalizable analysis

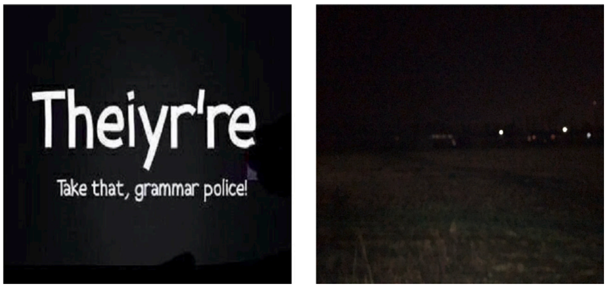
We conducted a Generalizable analysis of the proposed model on the Twitter dataset to investigate the model's generalization performance and explain the selection of feature extractors. In this section, different pre-trained models are employed to extract image and text data. There are three variants: ResNet-BERT, DenseNet-BERT, and DenseNet-RoBERTa. ResNet [70] and DenseNet [56] serve as image feature extractors. while BERT [49] and RoBERTa [58] as text feature extractor. The dimension of the image feature is 500, and the dimension of the text feature is 768. The model proposed in this paper is directly used with the pre-trained models, and the experimental results are shown in Fig. 3(b). As expected, our model can achieve better results when applying the pre-trained model with stronger performance.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. Attention visualization of two examples.



(a) (b)

Fig. 5. Examples for error analysis.

5.6. Case study and visualization

To further explore how the proposed model (SAHFN) learns sarcasm information, we present a case analysis and attention visualization on two samples whose labels are sarcasm and non-sarcasm, respectively. The two samples are from the Twitter dataset. These samples are multimodal, consisting of image, text, and attribute modalities. The sarcasm detection model proposed in this paper needs to use sentiment information to enhance its ability to extract representations from different modalities. It also needs to consider the inter-modal relationship. The visualization result of the attention is shown in Fig. 4. When the model employs multimodal data as input, it needs to capture the intra-modal information fully. As shown in Figs. 4(b) and 4(f), the model pays more attention to the local information of the image because of the contribution of attribute words. At the same time, after fusing sentiment vectors, the model can fully extract the text information and attribute information, as exhibited in the hot maps Figs. 4(c), 4(d), 4(g), 4(h). Consequently, our model becomes capable of learning and understanding more intra- and inter-modal relationships with the help of sentiment information and modal interaction to improve the ability of sarcasm detection.

5.7. Error analysis

Although the proposed model performs well in multimodal sarcasm detection tasks, there are still areas that could be enhanced. Upon observing the prediction results of the test set, we have found that certain words from the text, following sentence splitting, are not present in the emotion dictionary, which led to sparse sentiment vectors. Unfortunately, this situation limits the model's effective use of sentiment information. In addition, we have also observed that some images in the samples are simplistic or lack clarity, as exemplified by Figs. 5(a) and 5(b). In Fig. 5(a), only some words are on a black background. Furthermore, in Fig. 5(b), the scene is dark, which makes it difficult for our model to extract adequate visual information. To further enhance the performance of sarcasm detection, the utilization of more powerful image extraction technologies could be beneficial.

6. Conclusions and future work

Due to the weaknesses in current multimodal sarcasm detection, this paper presents a model based on the hierarchical fusion mechanism of sentiment perception to capture the association between different modalities, thus reducing redundant information. In addition, the low-level feature is also used as a guide to achieving the final modal fusion of feature vector weight calculation and then predicting labels. Simultaneously, contrastive losses are used to enhance the correlation between modalities. The experimental results demonstrate that the proposed model has certain advantages.

In future research, the proposed model will be evaluated on other public datasets to verify the model's generalization. In addition, we will also focus on how to capture more fine-grained information applied to sarcasm detection tasks. Furthermore, we are also prepared to extract the potential semantic information of images and texts in the way of image and text generation rather than focusing on modal fusion only.

Ethical approval

This article does not contain any studies with human participants or animals performed by any authors.

CRediT authorship contribution statement

Hao Liu: Conceptualization, Methodology, Software, Writing – original draft. **Runguo Wei:** Data curation, Visualization. **Geng Tu:** Investigation, Visualization. **Jiali Lin:** Writing – review & editing. **Cheng Liu:** Data curation. **Dazhi Jiang:** Supervision, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Natural Science Foundation of Guangdong Province (2024A1515010239), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence) (2019KZDZX1030), 2020Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D), Science and Technology Major Project of Guangdong Province (STKJ2021005, STKJ202209002, STKJ2023076), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology (2020B1212060078).

References

[1] R.W. Gibbs, On the psycholinguistics of sarcasm., J. Exp. Psychol.: Gen. 115 (1) (1986) 3.

[2] S. Dews, E. Winner, Muting the meaning a social function of irony, Metaphor Symb. 10 (1) (1995) 3–19.

[3] Q. Liu, X. Geng, Y. Wang, E. Cambria, D. Jiang, Disentangled retrieval and reasoning for implicit question answering, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12.

[4] D. Jiang, H. Liu, R. Wei, G. Tu, CSAT-FTCN: a fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition, Cogn. Comput. 15 (3) (2023) 1082–1091.

[5] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, D. Jiang, Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models, Knowl.-Based Syst. 235 (2022) 107598.

[6] D. Jiang, H. Liu, G. Tu, R. Wei, E. Cambria, Self-supervised utterance order prediction for emotion recognition in conversations, Neurocomputing (2024) 127370.

[7] G. Tu, R. Jing, B. Liang, M. Yang, K.-F. Wong, R. Xu, A training-free debiasing framework with counterfactual reasoning for conversational emotion detection, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15639–15650.

[8] I. Chaturvedi, E. Cambria, R.E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Inf. Fusion 44 (2018) 65–77.

[9] S. Dhelim, N. Aung, M.A. Bouras, H. Ning, E. Cambria, A survey on personality-aware recommendation systems, Artif. Intell. Rev. (2022) 1–46.

[10] M. Zhang, Y. Zhang, G. Fu, Tweet sarcasm detection using deep neural network, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2449–2460.

[11] T. Xiong, P. Zhang, H. Zhu, Y. Yang, Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2115–2124.

[12] Q. Lin, J. Liu, R. Mao, F. Xu, E. Cambria, TECHS: Temporal logical graph networks for explainable extrapolation reasoning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1281–1293.

[13] R. Schifanella, P. De Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 1136–1145.

[14] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion (2022).

[15] M. Bedi, S. Kumar, M.S. Akhtar, T. Chakraborty, Multi-modal sarcasm detection and humor classification in code-mixed conversations, IEEE Trans. Affect. Comput. (2021).

[16] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowleNet: Knowledge fusion network for multimodal sarcasm detection, Inf. Fusion 100 (2023) 101921.

[17] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, H.T. Shen, Cross-modal attention with semantic consistence for image–text matching, IEEE Trans. Neural Netw. Learn. Syst. 31 (12) (2020) 5412–5425.

[18] C. Wen, G. Jia, J. Yang, DIP: Dual incongruity perceiving network for sarcasm detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2540–2550.

[19] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2506–2515.

[20] A. Kumar, S.R. Sangwan, A. Arora, A. Nayyar, M. Abdel-Basset, et al., Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network, IEEE Access 7 (2019) 23319–23328.

[21] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, J. Li, Modeling incongruity between modalities for multimodal sarcasm detection, IEEE MultiMedia 28 (2) (2021) 86–95.

[22] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, R. Xu, Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 4707–4715.

[23] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of LREC 2022, 2022.

[24] P. Chaudhari, C. Chandankhede, Literature survey of sarcasm detection, in: 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET, IEEE, 2017, pp. 2041–2046.

[25] D. Maynard, M.A. Greenwood, Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14, ELRA, 2014, pp. 4238–4243.

[26] J. Cui, Z. Wang, S.-B. Ho, E. Cambria, Survey on sentiment analysis: evolution of research methods and topics, Artif. Intell. Rev. (2023) 1–42.

[27] A.G. Prasad, S. Sanjana, S.M. Bhat, B. Harish, Sentiment analysis for sarcasm detection on streaming short text data, in: 2017 2nd International Conference on Knowledge Engineering and Applications, ICKEA, IEEE, 2017, pp. 1–5.

[28] S. Mukherjee, P.K. Bala, Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering, Technol. Soc. 48 (2017) 19–27.

[29] A. Joshi, V. Sharma, P. Bhattacharyya, Harnessing context incongruity for sarcasm detection, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 757–762.

[30] D. Kovaz, R.J. Kreuz, M.A. Riordan, Distinguishing sarcasm from literal language: Evidence from books and blogging, Discourse Process. 50 (8) (2013) 598–615.

[31] A. Ghosh, T. Veale, Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 482–491.

[32] T. Young, V. Pandelea, S. Poria, E. Cambria, Dialogue systems with audio context, Neurocomputing 388 (2020) 102–109.

[33] Y. Ren, D. Ji, H. Ren, Context-augmented convolutional neural networks for twitter sarcasm detection, Neurocomputing 308 (2018) 1–7.

[34] Y. Tay, A.T. Luu, S.C. Hui, J. Su, Reasoning with sarcasm by reading in-between, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1010–1020.

[35] T. Ptáček, I. Habernal, J. Hong, Sarcasm detection on czech and english twitter, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 213–223.

[36] M. Khodak, N. Saunshi, K. Vodrahalli, A large self-annotated corpus for sarcasm, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 1–6, URL https://aclanthology.org/L18-1102.

[37] S. Zhang, X. Zhang, J. Chan, P. Rosso, Irony detection via sentiment-based transfer learning, Inf. Process. Manage. 56 (5) (2019) 1633–1644.

[38] P. Verma, N. Shukla, A. Shukla, Techniques of sarcasm detection: A review, in: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE, IEEE, 2021, pp. 968–972.

[39] S. Poria, E. Cambria, D. Hazarika, P. Vij, A deeper look into sarcastic tweets using deep convolutional neural networks, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1601–1612.

[40] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, R. Mihalcea, CASCADE: Contextual sarcasm detection in online discussion forums, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1837–1848.

[41] Y.A. Kolchinski, C. Potts, Representing social media users for sarcasm detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1115–1121.

[42] N. Babanejad, H. Davoudi, A. An, M. Papagelis, Affective and contextual embedding for sarcasm detection, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 225–243.

[43] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, S. Poria, Towards multimodal sarcasm detection (an _Obviously_Perfect paper), in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4619–4629.

[44] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, 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), 2018, pp. 2236–2246.

[45] R. Mao, K. He, X. Zhang, G. Chen, J. Ni, Z. Yang, E. Cambria, A survey on semantic processing techniques, Inf. Fusion 101 (2024) 101988.

[46] M. Malik, D. Tomás, P. Rosso, How challenging is multimodal irony detection? in: E. Métais, F. Meziane, V. Sugumaran, W. Manning, S. Reiff-Marganiec (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2023, pp. 18–32.

[47] X. Wang, X. Sun, T. Yang, H. Wang, Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data, in: Proceedings of the First International Workshop on Natural Language Processing beyond Text, 2020, pp. 19–29.

[48] A. Kamal, M. Abulaish, Cat-bigru: Convolution and attention with bi-directional gated recurrent unit for self-deprecating sarcasm detection, Cogn. Comput. 14 (1) (2022) 91–109.

[49] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 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), 2019, pp. 4171–4186.

[50] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, R. Schifanella, Transformer-based models for multimodal irony detection, J. Ambient Intell. Humaniz. Comput. 14 (6) (2023) 7399–7410.

[51] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 1767–1777.

[52] S. Pramanick, A. Roy, V.M. Patel, Multimodal learning using optimal transport for sarcasm and humor detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3930–3940.

[53] H. Liu, B. Yang, Z. Yu, A multi-view interactive approach for multimodal sarcasm detection in social internet of things with knowledge enhancement, Appl. Sci. 14 (5) (2024) 2146.

[54] H. Fu, H. Liu, H. Wang, L. Xu, J. Lin, D. Jiang, Multi-modal sarcasm detection with sentiment word embedding, Electronics 13 (5) (2024) 855.

[55] Y. Zhang, Y. Liu, Q. Li, P. Tiwari, B. Wang, Y. Li, H.M. Pandey, P. Zhang, D. Song, CFN: a complex-valued fuzzy network for sarcasm detection in conversations, IEEE Trans. Fuzzy Syst. 29 (12) (2021) 3696–3710.

[56] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[57] D. Ghosh, W. Guo, S. Muresan, Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1003–1012.

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019.

[59] S. Bird, NLTK: the natural language toolkit, in: Proceedings of the COLING/ACL on Interactive Presentation Sessions, 2006, pp. 69–72.

[60] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.

[61] J. Ni, R. Mao, Z. Yang, H. Lei, E. Cambria, Finding the pillars of strength for multi-head attention, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 14526–14540, URL https://aclanthology.org/2023.acl-long.812.

[62] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.

[63] H. Liu, W. Wang, H. Li, Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4995–5006, URL https://aclanthology.org/2022.emnlp-main.333.

[64] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751, URL https://aclanthology.org/D14-1181.

[65] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in Twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2506–2515, URL https://aclanthology.org/P19-1239.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16 × 16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, pp. 1–22, URL https://openreview.net/forum?id=YicbFdNTTy.

[67] N. Xu, Z. Zeng, W. Mao, 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, 2020, pp. 3777–3786.

[68] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1383–1392.

[69] S. Gupta, A. Shah, M. Shah, L. Syiemlieh, C. Maurya, FiLMing multimodal sarcasm detection with attention, in: International Conference on Neural Information Processing, Springer, 2021, pp. 178–186.

[70] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.