Multimodal Dual Perception Fusion Framework for Multimodal Affective Analysis

Multimodal Dual Perception Fusion Framework for Multimodal Affective Analysis

Qiang Lu[a], Xia Sun[a,*], Yunfei Long[b], Xiaodi Zhao[a], Wang Zou[a], Jun Feng[a,*], Xuxin Wang[a]

[a]*School of Information Science and Technology, Northwest University, Xi'an 710127, China*
[b]*School of Computer Science and Electrical Engineering, University of Essex, Colchester CO43SQ, UK*

Abstract

The misuse of social platforms and the difficulty in regulating post contents have culminated in a surge of negative sentiments, sarcasms, and the rampant spread of fake news. In response, Multimodal sentiment analysis, sarcasm detection and fake news detection based on image and text have attracted considerable attention recently. Due to that these areas share semantic and sentiment features and confront related fusion challenges in deciphering complex human expressions across different modalities, integrating these multimodal classification tasks that share commonalities across different scenarios into a unified framework is expected to simplify research in sentiment analysis, and enhance the effectiveness of classification tasks involving both semantic and sentiment modeling. Therefore, we consider integral components of a broader spectrum of research known as multimodal affective analysis towards semantics and sentiment, and propose a novel multimodal dual perception fusion framework (MDPF). Specifically, MDPF contains three core procedures: 1) Generating bootstrapping language-image Knowledge to enrich origin modality space, and utilizing cross-modal contrastive learning for aligning text and image modalities to understand underlying semantics and interactions. 2) Designing dynamic connective mechanism to adaptively match image-text pairs and jointly employing gaussian-weighted distribution to intensify semantic sequences. 3)

*Corresponding author
Email addresses: nwulq@stumail.nwu.edu.cn (Qiang Lu), raindy@nwu.edu.cn (Xia Sun), yl20051@essex.ac.uk (Yunfei Long), zhaoxiaodi@stumail.nwu.edu.cn (Xiaodi Zhao), zouwang@huat.edu.cn (Wang Zou), fengjun@nwu.edu.cn (Jun Feng), 13289381224@163.com (Xuxin Wang)

Constructing a cross-modal graph to preserve the structured information of both image and text data and share information between modalities, while introducing sentiment knowledge to refine the edge weights of the graph to capture cross-modal sentiment interaction. We evaluate MDPF on three publicly available datasets across three tasks, and the empirical results demonstrate the superiority of our proposed model.
Keywords: multimodal sentiment analysis, sarcasm detection, fake news detection, multimodal affective analysis, multimodal dual perception fusion

1. Introduction

Social data has evolved from being predominantly text-based to a rich, diversified format where images and text coexist harmoniously. This transformation has led individuals to increasingly rely on multimodal data to articulate their affective viewpoints. However, the misuse of these platforms and the inherent challenges in regulating posted content have culminated in a surge of negative sentiments, sarcasm, and the rampant spread of fake news [27, 36, 41, 60]. In response, the fields of multimodal sentiment analysis (MSA), multimodal sarcasm detection (MSD), and multimodal fake news detection (MFND) have emerged, focusing on the intertwined analyses of images and text to address these concerns.

Multimodal sentiment analysis, sarcasm detection, and fake news detection all fall within the realm of human affective behavior [1, 4, 19, 58], and these areas are integral components of a broader spectrum of research known as multimodal affective analysis, involving same emotional mechanisms, namely, dual modeling of semantics and sentiment. As illustrated in Figure 1, semantics represents the factual content described by text and images, and sentiment represents the emotional tendencies contained in text and images. The orderly fusion of factual content and sentiment tendencies can achieve the mutual complementation of semantic and emotional information in cases where the text and image are consistent, so as to accurately predict the sentiment polarity, as shown in Figure 1 (a). Figure (a) lacks explicit sentiment expression, and the discrimination of its sentiment polarity requires the combination of semantic content with the positive sentiment elements conveyed through the azure sea and burning sun in the image. In

Task 1：Sentiment Analysis	Task 2：Sarcasm Detection	Task 3：Fake News Detection
Semantics： holiday, beach (Text) sun, beach, sea (Image)	Semantics: warm, dry conditions (Text) snow, wet ground, cloudy (Image)	Semantics: little Syrian girl, chewing gum (Text) neatly dressed, smiling (Iamge)
Sentiment: Neutral (Text) Positive (Image)	Sentiment: Positive (Text) Negative (Image)	Sentiment: Negative (Text) Positive (Image)
Prediction：Positive	Prediction：Sarcasm	Prediction：Fake
(a) We spent our holiday at the beach.	(b) Gotta love the warm and dry conditions !	(c) Little Syrian girl sells chewing gum on the street so she can feed herself.

Figure 1: Several examples from MSA, MSD, and MFND. (a) lacks explicit sentiment expression which need to be complemented by the sentiment conveyed through image. (b) distinguishes sarcasm through the significant differences between sentiment word and semantics of the image. (c) illustrate the impact of semantics and sentiment information on fake news detection.

incongruent scenarios, the mutual comparison of semantics and sentiment can help to judge whether it is ironic or false, as shown in Figure 1 (b) sarcasm detection and Figure 1 (c) fake news detection. In Figure (b), the text descriptions of *warm* and *dry* form a strong semantic contrast with the image object *snow*, while the word *like* creates a sentiment contrast with the cloudy area, which further exacerbates the degree of inconsistency. Fake news in Figure (c) narrates a war-like situation, while the image implicitly depicts a happy sentiment. Integrating these multimodal classification tasks with commonalities in different scenarios into a unified multimodal framework is expected to simplify the research work of multimodal sentiment analysis and improve the effectiveness of multimodal classification tasks involving dual semantic and sentiment modeling.

In addressing the complexities of multimodal affective content interpretation, we

integrate multimodal sentiment analysis, multimodal sarcasm detection, and multimodal fake news detection into a unified approach within the broader realm of multimodal affective analysis towards semantics and sentiment. This integration acknowledges the shared features and challenges across these domains, leading us to propose the Multimodal Dual Perception Fusion Framework (MDPF). Distinguished from traditional sentiment classifications, MDPF focuses on a broader spectrum of multimodal affective classification by combining semantic and emotional cues for comprehensive multimodal analysis. Specifically, we enrich multimodal representations with image-derived knowledge by leveraging a bootstrapping language-image pre-trained model, and employ cross-modal contrastive learning for precise text-image alignment, capturing the nuances of multimodal expressions. Meanwhile, MDPF applies dynamic routing attention mechanisms and probability density functions to capture shared and private features to identify and enhance semantics across modalities. Additionally, it integrates sentiment knowledge to refine cross-modal interactions by performing graph convolution, culminating in the fusion of semantic intensified and sentiment interactive sequences. This novel approach marks a significant advancement in affective classification, enhancing the interpretation of complex affective phenomena across textual and visual data, and providing new research ideas for multimodal sentiment analysis research. The main contributions of this work are summarized as follows:

- We introduce the Multimodal Dual Perception Fusion Framework (MDPF), integrating multimodal classification tasks involving dual modeling of semantic and sentiment in different application scenarios into a unified framework. This approach not only promises to streamline sentiment analysis research but also enhances the effectiveness of detecting and analyzing complex affective phenomena in social media content.
- By infusing the feature encoding process with image description knowledge and cross-modal contrastive learning, our approach significantly enriches the multimodal sentiment space, ensuring a robust alignment of textual and visual modalities with a sentiment focus.
- We develop a dynamic connective mechanism alongside a gaussian-weighted

distribution approach, adept at distilling shared and private sentiment features, and incorporate sentiment knowledge into the cross-modal graph. This innovation significantly enhances our proposed framework's ability to navigate and interpret complex affective landscapes effectively.

The remainder of this paper is organized as follows. After introducing related works in Section 2, we propose a Multimodal Dual Perception Fusion Framework in Section 3. Then, the experimental details and analysis are described in Section 4. Finally, we summarize our work and provide a direction for future work in Section 5.

2. Related work

Multimodal fusion methods aim to construct unified semantic space combining image and text information to perform multimodal task [10, 20, 46]. However, discrepancies across different modalities hinder the development of models. Therefore, various multimodal classification models such as multimodal sentiment analysis, sarcasm detection, and fake news detection design advanced fusion strategies to address this issue.

2.1. Multimodal sentiment analysis

MSA utilizes multimodal fusion methods to map multimodal features into a unified semantic space for multimodal sentiment classification, and it focuses on modeling modality interaction between image and text. Xu et al. [49] proposed a co-memory network to integrate visual contents and textual words via an interaction strategy. The above method may cause information loss, Jiang et al. [14] designed an interactive information fusion mechanism to interactively learn cross-modal representations. Yang et al. [52] relied on the interaction of text and image modalities to fine-tune the pre-trained BERT model. Rahman et al. [34] allowed BERT and XLNet to accept multimodal non-linguistic data during fine-tuning for multimodal sentiment analysis. Zhu et al. [57] believed that the correspondence between image regions and words in image-text pairs facilitates cross-modal interaction. Additionally, Yadav et al. [51] used deep multi-level attention to exploit the correlation between image and text modalities to improve multimodal sentiment analysis.

2.2. Multimodal sarcasm detection

MSD is a subtask of MSA that aims to identify whether the text description differs from the image content. Previous studies have attempted to model sarcasm via fusion and interaction across modalities. Schifanella et al. [35] combined visual semantics with text features from an external dataset, while the other used pre-trained visual neural networks to analyze multimodal sarcasm. Cai et al. [3] tried to use multimodal hierarchical fusion to detect multimodal sarcasm. To address the issue in reasoning with multimodal sarcasm, Xu et al. [50] designed a decomposition and relation network based on cross-modal contrast and semantic association for sarcasm detection.

However, previous methods have overlooked the deep implicit sarcasm relationship. Therefore, recent studies attempt to model modality incongruity to fusion image and text information. Liang et al. [22] constructed cross-modal graphs to model the incongruity of image and text. To address the issue that models lack the flexibility to diverse image-text pairs, Tian et al. [39] employed dynamic paths and hierarchical co-attention adapting to model cross-modal incongruity. Wen et al. [45] argued the challenge of MSD from implicit intention and intrinsic conflict, and proposed a semantic intensified distribution and siamese sentiment Contrastive Learning method. Jia et al. [13] believed the spurious correlations can significantly hinder the generalization capability, and they proposed a debiasing multimodal sarcasm detection framework to alleviate the harmful effects of biased textual factors on robust OOD generalization.

2.3. Multimodal fake news detection

MFND task based on contents is similar to MSD which aims to identify fake news as true or false through the image-text pair. Early studies leverage cross-modal discriminative patterns to integrate image and text to improve the accuracy. Khattar et al. [16] used a bimodal variational autoencoder combined with a binary classifier to perform fake news detection tasks. To address the specific features could not be transferred to newly emerging events, Zhang et al. [55] adopted multimodal knowledge-aware networks and event memory networks as building blocks for social media fake news detection. In view of the fact that many existing methods ignore the correlation between text and visual features and lead to suboptimal results, Wei et al. [44] proposed a novel

cross-modal knowledge distillation function as a soft target to guide unimodal network modeling and train a fusion model for multimodal fake news detection.

Recently, some methods have worked on the principles of cross-modal incongruity. Chen et al. [6] calculated Kullback-Leibler (KL) divergence to measure cross-modal consistency. To improve the performance of the decision-making processing, Wang et al. [42] proposed a cross-modal contrastive learning method for multimodal fake news. Moreover, Chen et al. [7] proposed a causal intervention and counterfactual reasoning-based debiasing framework to mitigate multimodal biases from a causality perspective. Wu et al. [46] adopted multimodal fusion and inconsistency reasoning to discover multimodal inconsistent semantics for interpretable fake news detection. Additionally, Peng et al. [31] proposed contextual semantic representation learning to introduce contextual information into the representation learning process for multimodal fake news detection.

Despite multimodal fusion methods have achieved promising results, existing approaches only focus on individual tasks and have not yet established the inherent connections among these tasks. Given that MSA, MSD, and MFND fall within the realm of human affective behavior and share common fundamental characteristics. Meanwhile, these tasks confront related challenges in deciphering complex human expressions across different modalities. Therefore, we aim to design a multimodal dual perception fusion framework to integrate these tasks into a unified multimodal model.

3. Methodology

In this work, we propose a novel Multimodal Dual Perception Fusion Framework (MDPF) for multimodal sentiment analysis, sarcasm detection, and fake news detection. The proposed method defines the three tasks as constructing a fused sequence of semantics and sentiment. Specifically, we first utilize Bootstrapping Language-Image Pre-training to generate image descriptions as external knowledge and combine image-text contrastive learning for cross-modal alignment. Then, we design a dual perception mechanism that contains a semantic joint perception module and a sentiment interactive perception module. The semantic joint perception module utilizes a dynamic connective

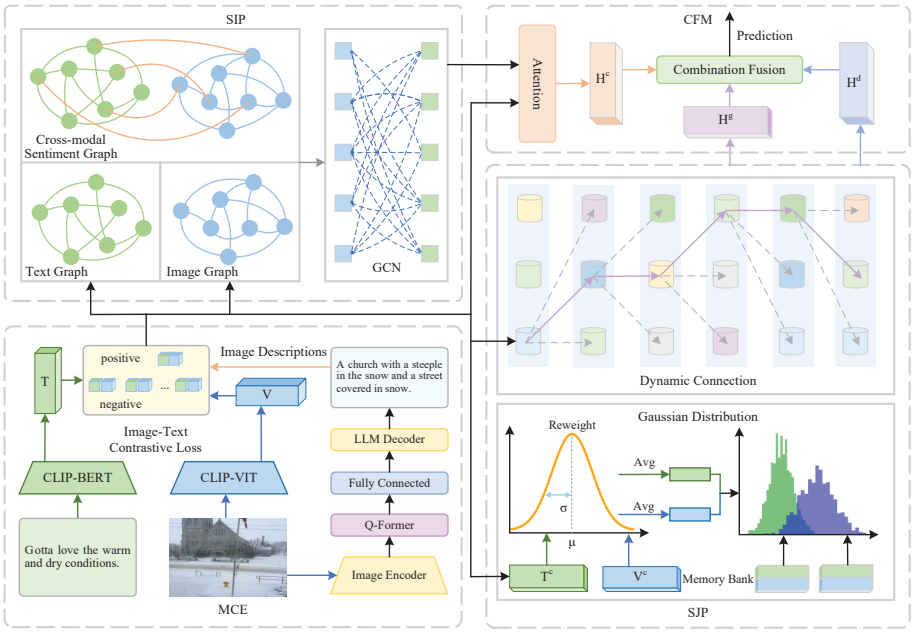

Figure 2: Overall architecture of our proposed MDPF model. MCE, SJP, SIP, and CFM denote the multimodal contrastive encoding, semantic joint perception, sentiment interactive perception, and combination fusion module, respectively.

mechanism and gaussian- weighted distribution to capture shared and private features for constructing semantic intensified sequences. Meanwhile, the sentiment interactive perception module integrates textual and visual graphs to build a cross-modal graph convolutional network for generating sentiment interactive sequences. Finally, a cross-modal combination fusion module is employed to merge the dual perceptual information for multimodal classification. The framework of the proposed MDPF is shown in Figure 2. Next, we will provide the task and notation definitions and elaborate on each specific technique in detail.

3.1. Task and Notation Definition

Multimodal sentiment analysis, sarcasm detection and fake news detection involving both text and images are integral components of a broader spectrum of research known as multimodal affective analysis that can be defined as: given a collection of N training

samples denoted as $S = \{s_1, s_2, ..., s_N\}$, each sample $s = \{x_i^v, x_i^t\}$ contains an image-text pair which the image is split into m patches and text contains n words is denoted as $Image = \{e_i^v \,|\, 1 \leq j \leq m\}$ and $Text = \{e_j^t \,|\, 1 \leq i \leq n\}$, where e_i^v denotes the i-th patch of the image, and e_j^t denotes the j-th word of the sentence. The goal of task is to design a classifier $\mathcal{F}$ which jointly utilizes image and text modalities to predict the truth label $\hat{y}_i = \mathcal{F}(x^v, x^t, \theta)$ for each image-text pair.

3.2. Multimodal Contrastive Encoding

The image and text information are first input into feature encoders. The image and text embedding are obtained by utilizing a ViT [8] and Transformer [40] model from CLIP [33] to enhance the cross-modal alignment and interaction capabilities within the original modality space. The process is formulated as follows:

$$\begin{aligned} V =& [v_1, v_2, ..., v_m] = CLIP\_ViT(Image) \\ T =& [t_1, t_2, ..., t_n] = CLIP\_Transformer(Text) \end{aligned} \tag{1}$$

where $V \in \mathbb{R}^{m \times d_v}$ and $T \in \mathbb{R}^{n \times d_t}$ represent the output embeddings of image and text. $v_i \in \mathbb{R}^{d_v}$ denotes the i-th patch of image, and $t_j \in \mathbb{R}^{d_t}$ denotes the j-th word. d_v, d_t denote the dimension of image and text embedding.

3.2.1. Bootstrapping Language-Image Knowledge Construction

The limited richness of modal information representation is caused by low-level image features and the absence of direct supervision for image-text alignment in the multimodal task. Therefore, we construct external knowledge by generating image descriptions D via Bootstrapping Language-Image pre-training model [18] to enhance the multimodal representation.

$$D = LLM(FC(QFormer(x^v))) \tag{2}$$

Knowledge construction is pre-trained in two stages: In the first stage, the Q-Former learns visual representations that are most relevant to the text. In the second stage, visual-language generative learning is conducted by connecting the output of the Q-Former

to a frozen Large Language Model (i.e., OPT [56]), allowing the visual representations produced by the Q-Former to be directly interpreted by the Large Language Models(LLMs).

Based on image descriptions, we restructure the encoding stage to introduce external knowledge. Specifically, we continued to use ViT and BERT trained based on CLIP to encode the image and the text concatenated with image description, respectively.

$$\begin{aligned} V^c &= CLIP\_ViT(V) \\ T^c &= CLIP\_Transformer(concat(T, D)) \end{aligned} \tag{3}$$

Where $V^c \in \mathbb{R}^{m \times d_v}$ represents the output embedding of images, and $T^c \in \mathbb{R}^{n \times d_t}$ is the output embedding representing the combination of text and image description.

3.2.2. Cross-modal Contrastive learning

After obtaining the output embeddings, we align the image and text modalities to comprehensively understand the underlying semantics and interaction between the text and image. We adopt the InfoNCE loss [29] to construct cross-modal contrastive loss. For each image text pair, image-to-text contrastive loss is defined as:

$$\mathcal{L}^{v2t} = -\log \frac{\exp(\text{sim}(v_i^c, t_i^c)/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(v_i^c, t_j^c)/\tau)} \tag{4}$$

Where v_i^c, t_i^c are the layer normalized representations from V^c, T^c. sim denotes similarity measured by the dot product of the image-text pair. τ represents the learnable temperature parameters. Then, the text-to-image contrastive loss $\mathcal{L}^{t2v}$ is constructed in a similar way.

$$\mathcal{L}^{t2v} = -\log \frac{\exp(\text{sim}(t_i^c, v_i^c)/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(t_i^c, v_j^c)/\tau)} \tag{5}$$

Where the contrastive learning generates sets of positive and negative instances via a batch of N input pairs (v_a^a, t_b^a). Specifically, there are one positive instance and $N-1$ negative instances for each pair in the batch. The positive sample is formed by the most similar image text pairs in the batch, represented as $(v_i^a, t_j^a)_{i=j}$, while other $N-1$ image text pairs are defined as negative samples, represented as $(v_i^a, t_j^a)_{i \neq j}$. Finally, the optimization objective of cross-modal contrastive learning is defined as:

$$\mathcal{L}_{cl} = \frac{1}{2N}(\mathcal{L}^{v2t} + \mathcal{L}^{t2v}) \tag{6}$$

$\mathcal{L}_{cl}$ obtains the semantic alignment unimodal feature v^a, t^a by maximizing the similarity between positive examples and minimizing the similarity between negative examples. After completing the contrastive encoding, aligned image-text semantics are input into dual perception module to construct semantic and sentiment sequence.

3.3. Semantic Joint Perception

Given the presence of shared and private features in multimodal data, shared features manifest as semantic congruity to enhance semantics. In contrast, private features display semantic incongruity to distinguish and complement semantics [26, 47]. Therefore, we design a semantic joint perception module with a dynamic connection mechanism and Gaussian-weighted distribution, which captures shared and private features in parallel to construct semantic intensified sequences, speeding up the calculation while reducing the introduction of noise and redundant information, thereby achieving cross-modal semantic enhancement and supplementation.

3.3.1. Dynamic connective mechanism

Shared features enhance semantics by matching images and text combinations that jointly describe the same entity. Therefore, we calculate dynamic connective weight using a routing attention mechanism [39] to capture the matching probability between each image patch and word token. Firstly, we introduce a graph attention mechanism that aggregates text and graph edge relationships to enhance node representation.

$$\begin{aligned} T^a &= \frac{\exp(\sigma(a^t[t_i^c \cdot W_T || t_j^c \cdot W_T]))}{\sum_{i=1}^{m} \exp(\sigma(a^t[t_i^c \cdot W_T || t_k^c \cdot W_T]))} * T^a \\ V^a &= \frac{\exp(\sigma(a^v[v_i^c \cdot W_V || v_j^c \cdot W_V]))}{\sum_{i=1}^{n} \exp(\sigma(a^v[v_i^c \cdot W_V || v_k^c \cdot W_V]))} * V^a \end{aligned} \tag{7}$$

Where σ denotes the $LeakyReLU$ activation function. $a^t \in \mathbb{R}^{2d_t}, a^v \in \mathbb{R}^{2d_v}$ and $W_T \in \mathbb{R}^{d_t \times d_t}, W_V \in \mathbb{R}^{d_v \times d_v}$ are the learnable parameter. Then, the dynamic routing embedding $route_i$ can be calculated as follows:

$$route_i = softmax(\frac{Q_T(K_V)^{\mathrm{T}}}{\sqrt{d_k}} \otimes \sum_{i=1}^{k} \alpha_i A_i) V_V \tag{8}$$

Where $route_i \in \mathbb{R}^{h \times d_h}$ is the dynamic routing embedding, and $d_t = n \cdot d_h$. $Q_T = t^a w_q \in \mathbb{R}^{n \times d_r}$, $K_V = v^a w_k \in \mathbb{R}^{m \times d_r}$, $V_V = v^a w_v \in \mathbb{R}^{m \times d_r}$ denote the query, key

and value representations, and $w_q \in \mathbb{R}^{d_t \times d_r}, w_k \in \mathbb{R}^{d_v \times d_r}, w_v \in \mathbb{R}^{d_v \times d_r}$ are learned parameter. $\alpha_i = softmax(MLP(Att\_Pool(v^i)))$ denotes the routing matching probability, and Att_Pool denotes the adaptive average pooling function. $A_i \in \mathbb{R}^{n \times m}$ is the matching coefficients between each image-text pair. If the image patch is within the attention span of the text target, the value of coefficients is set to 1, otherwise set to 0. Then, we treat the $route_i$ as a attention function head, executing h heads in parallel across the hidden dimensions via multi-head self-attention.

$$H^r = Concat(head_1, head_2, ..., head_h)w_o \tag{9}$$

Where $H^r \in \mathbb{R}^{n \times d_h}$ represents the dynamic connective representation. $Concat$ denotes the concatenation operation. $w_0 \in \mathbb{R}^{d_t \times d_t}$ is the weight matrix. Finally, H^r is input into a dynamic connective layer to capture the shared features.

$$\begin{aligned} H^m_{k-1} &= LayerNorm(MHA(H^r) + H^r) \\ H^d_k &= LayerNorm(FNN(H^m_{k-1}) + H^m_{k-1}) \end{aligned} \tag{10}$$

Where $H^d \in \mathbb{R}^{n \times d}$ is the output embedding of $k-th$ layer which contains the shared features, among $d_v = d_t = d$. MHA denotes the multi-head self-attention and FNN denotes the feed-forward network. The dynamic connective mechanism focuses on different subspaces of the image and text while leveraging routing matching matrices and attention mechanisms to establish more effective information transmission and interaction across various modalities.

3.3.2. Gaussian-weighted distribution

Private semantics are significantly greater than shared semantics in terms of mean and variance. Therefore, we adopt a gaussian-weighted distribution to model private semantics. First, we design a reweighting strategy to find the most relevant representations of private semantics.

$$v^r = \sigma((v^a \cdot (t^a)^T) * e^\tau) \cdot v^a \tag{11}$$

Where $v^r \in \mathbb{R}^{m \times d_v}$ is the reweighted image embedding, and e^τ is the temperature parameter. The reweighted text $t^r \in \mathbb{R}^{n \times d_t}$ is designed in a similar way. Then, we

calculate the mean and variance of private samples through a gaussian distribution and maintain the gaussian distribution with a memory bank.

$$\mu = \sum_{i=1}^{N} \mathrm{M}(v_i^r, t_i^r) \tag{12}$$

$$\sigma = \sqrt{\sum_{i=1}^{N} (\mathrm{M}(v_i^r, t_i^r) - \mu)^2} \tag{13}$$

Where μ and σ are the mean and variance values. $M(v_i^r, t_i^r)$ is the maintained memory bank [12, 45], which stores the features of image-text pairs in a discrete storage unit and computes the cosine similarity between instances directly based on features in a non-parametric manner. After that, we use the probability density function to calculate the probability of samples falling within the private semantic region.

$$p^g = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\tau(\frac{\mathrm{M}(v_i^r, t_i^r) - \mu}{\sigma})^2} \tag{14}$$

The probability density function p^g is optimized using gradient descent during each learning iteration, and calculates the difference value between p^g and the original image-text pair as $\lambda = p^g - p$ to guide the modeling of private semantics. Finally, the gaussian-optimized image-text pairs v^g, t^g is integrated into the cross-modal attention mechanism to capture private features $H^g \in \mathbb{R}^{n \times d}$.

3.4. Sentiment Interactive Perception

The sentiment relations between image and text modalities are crucial for multimodal fusion in sentiment analysis, sarcasm detection, and fake news detection [25]. Image data contains rich spatial structural information, while text data encompasses semantic and syntactic dependency information. Modeling image-text pairs using cross-modal graph convolutional neural networks can preserve the structured information of both image and text data and share information between modalities, thereby better capturing cross-modal sentiment relations. Thus, we design a cross-modal graph convolutional network to construct sentiment interactive sequences. First, we generate textual graphs

based on SenticNet [5] and dependency tree.

$$D_{i,j}^{t} = \begin{cases} 1 + \delta(t_i^a), & \text{if } i = j \text{ or } \{t_i^c, t_j^c\} in \in T \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

Where $\delta(t_i^a) \in [-1, 1]$ represents sentiment value of i-th word retrieved from SenticNet. T is the origin dependency tree[1]. The visual graphs $D_{i,j}^{v}$ are generated in the same manner. Notably, the construction of visual edge weights is determined by judging whether different image patches belong to the same object:

$$D_{i,j}^{v} = \begin{cases} 1 + \delta(v_i^a), & \text{if } i = j \text{ or } \{t_i^c, t_j^c\} \in R \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Where $D_{i,j}^{v} \in \mathbb{R}^{n \times n}$ is the adjacency matrix of visual modality, and v_i^c, v_j^c denote the image patches. Due to the weights of the edges are important in graph information aggregation [22, 25], we construct edge weights for cross-modal graphs by measuring the distance between textual words and image objects, and introduce sentiment knowledge to calculate the sentiment score within each image-text pair to refine the edge weights.

$$D_{i,j}^{cross} = \begin{cases} 1 + \delta(t_i^a), & \text{if } D_{i,j}^{t} \text{ and } i < m, j < n \\ 1 + \lambda sim(v_i^a, t_i^a), & \text{if } i < m, j \geq n \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Where $D_{i,j}^{cross} \in \mathbb{R}^{(m+n)\times(m+n)}$ represents the cross-modal graph matrix. $sim(v_i^a, t_i^a)$ denotes the cosine similarity between textual words and image objects. $\lambda = e^{-\delta(t_i^a)\delta(a_j^a)}$. $|\delta(t_i^a) - \delta(a_j^a)|$ is the sentiment score to modulate the sentiment relations to refine edge weights by introducing the sentiment knowledge. $\delta(t_i^c)$ and $\delta(a_j^a)$ indicate the sentiment values of the dependency words and image patch attribute retrieved from SenticNet. If words and attribute are not present in the SenticNet, we set them to 0. Simultaneously, following the settings of [17], we configure the cross-modal graph as an undirected graph and set the diagonal nodes to 1.

[1]We use spaCy toolkit to construct the dependency tree: `https://spacy.io/`

Finally, we feed the cross-modal graph matrix $D_{i,j}^{cross}$ into graph convolutional network (GCN) and utilize a retrieval-based attention to generate sentiment interactive representation.

$$g_i^l = ReLU(\sum_{j=1}^{m+n} D_{i,j}^{cross} W_l \, g_j^{l-1} + b_l) \tag{18}$$

$$H^c = \sum_{t=1}^{m+n} \eta_t \cdot h_t^l, \;\; \eta_t = \frac{\exp(\beta_t)}{\sum_{i=1}^{m+n} \exp(\beta_i)}, \;\; \beta_t = \sum_{i=1}^{m+n} g_i^l \cdot h_t^l \tag{19}$$

Where g_i^l is the graph hidden representation, and $h_t^l = [H^t, H^v] = \{h_1^t, h_2^t, ..., h_n^t, h_1^v, h_2^v, ..., h_m^v\}$. H^t and H^v represent the textual and visual graph representations evolved from $D_{i,j}^t$ and $D_{i,j}^v$. β_t and η_t is the retrieval-based attention score and weight. $H^c \in \mathbb{R}^d$ is the final sentiment interactive representation.

3.5. Combination Fusion Module

Based on semantic intensified and sentiment interactive sequences, we project the shared, private and sentiment interactive representation into a cross-modal combination fusion module to achieve the multimodal classification.

$$\begin{aligned} M^d, \;\; M^g, \;\; M^c &= Mean(H^d, \;\; H^g, \;\; H^c) \\ y^f &= LN(W_f(M^d + M^g + M^c)) + b_f \\ \hat{y} &= softmax(W_y \cdot y^f + b_y) \end{aligned} \tag{20}$$

Where M_d, M_g and M_c are processed by the average function. $y^f \in \mathbb{R}^d$ is the multimodal fusion embedding. W_f, W_y, b_f, and b_y are the learned parameters. $\hat{y} \in \mathbb{R}^{d_p}$ is the predicted probability, and d_p is the dimensionality of possible labels.

3.6. Training Objective

The training of MDPF is to optimize all the parameters, and minimize the loss function via the standard gradient descent function. The overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{gce} + \lambda_1 \mathcal{L}_{cl} + \lambda_2 \mathcal{L}_{gs} \tag{21}$$

Where λ_1 and λ_2 are trade-off factors between different objective functions. $\mathcal{L}_{gce}$ is the global cross-entropy loss, $\mathcal{L}_{cl}$ is the cross-modal contrastive loss and $\mathcal{L}_{gs}$ is the gaussian-optimized loss.

4. Experiments

In this section, we initially describe the experimental datasets in Section 4.1, followed by the implementation details and baseline models in Sections 4.2 and 4.3. Then, we compare proposed MDPF with advanced baselines on MVSA, MSD and Twitter datasets in Section 4.4 to evaluate the performance. Next, we conduct an ablation study to analyze the contribution of MCE, SJP, SIP and CFM modules in Section 4.5, and introduce the parameter analysis in Section 4.6. We also introduce a case study to explain how bootstrapping language-image knowledge affects the proposed MDPF model in Section 4.7. Finally, we provide a visualization to discuss how the semantic joint perception and cross-modal sentiment interactive graphs assist the dual perception module in capturing semantic and sentiment information in Section 4.8.

4.1. Experimental datasets

We evaluate MDPF on three publicly available multimodal affective classification benchmarks: MVSA [28] for sentiment analysis, MSD [3] for multimodal sarcasm detection, and a multimodal fake news detection benchmark from Twitter [2].

MVSA comprises two datasets: MVSA-Single and MVSA-Multiple, tailored for sentiment analysis. MVSA-Single consists of 4,869 image-text pairs, excluding pairs with inconsistent image and text labels, while MVSA-Multiple encompasses 16,779 image-text pairs.

MSD contains 24,635 image-text pairs from Twitter that collects the specific hashtag (e.g., #sarcasm, etc.) as sarcastic examples, and collects pairs without such hashtags as non-sarcastic examples. To improve the quality, Cai et al. [3] discards tweets containing $sarcasm, sarcastic, irony, ironic$ as regular words and discards $URLs$.

Twitter is released for Verifying Multimedia Use task at MediaEval [2]. This database is created around widely known 11 real-world events, and it is divided into 9,596 fake tweets and 6,225 real tweets.

[2] http://multimediaeval.org

4.2. Implementation Details

We follow the previous studies [3, 15, 28] to process the datasets to ensure a fair comparison. We utilize the origin ViT and BERT, and pre-trained CLIP ViT and BERT [33] with 768 dimensions to initialize the image and text embeddings. The batch size is set to 32, and Adam is adopted as the optimizer with a learning rate of 0.00001. To optimize the model training, the epoch is set to 40 runs with random initialization, and the value of early-stopping is set to 5. The proposed MDPF and all baselines are implemented with PyTorch and performed by NVIDIA GeForce RTX 3090.

4.3. Baselines

We compare the performance of MDPF with strong baselines across three tasks, including MSA, MSS and MFND which are listed in Table 1, 2 and 3. The MSA baselines combine images and texts information for sentiment analysis are as follows:

- **CoMN** [49] exploits the correlation between image and text modalities to improve multimodal learning.
- **FENet** [14] uses an interactive information fusion mechanism to learn vision-specific textual representations and text-specific visual representations.
- **CM-BERT** [52] relies on the interaction of text and image modalities to fine-tune the pre-trained BERT model.
- **MAG** [34] allows BERT and XLNet to accept multimodal non-linguistic data during fine-tuning.
- **MVAN** [53] utilizes a continuously updated memory network to obtain deep semantic features of image text.
- **ITIN** [57] introduces a cross-modal alignment module to capture region-word correspondence, and on this basis fuses multimodal features through an adaptive cross-modal gating module.
- **DMLANet** [51] uses deep multi-level attention to exploit the correlation between image and text modalities to improve multimodal learning.

The MSD baselines combine images and texts information as follows:

- **HFM** [3] proposes a multimodal hierarchical fusion model to detect sarcasm.
- **Res-BERT** [30] uses BERT to encode text and combines text and image features for sarcasm prediction.
- **Att-BERT** [30] builds an attention mechanism to construct inter-modal attention to capture inter-modal inconsistencies.
- **InCrossMGs** [21] detects sarcasm by designing heterogeneous intra-modal and cross-modal graphs through graph convolutional networks.
- **CMGCN** [22] builds a cross-modal graph convolutional network to draw sarcasm-sarcasm relations.
- **HKEmodel** [23] Combines atomic-level consistency and atomic-level consistency based on graph convolutional networks to detect sarcasm.
- **MILNet** [32] designs local semantic guidance and global inconsistency learning modules for sarcasm detection.
- **DIP** [45] proposes a dual incongruity-aware network that uses Gaussian distribution and contrastive learning for sarcasm detection.
- **DynRT** [39] uses hierarchical joint attention to build dynamic paths to detect cross-modal sarcasm incongruity.
- **KnowleNet** [54] incorporates prior knowledge through ConceptNet knowledge and captures cross-modal semantic similarity for sarcasm prediction.
- **SAHFN** [24] introduces a hierarchical fusion model to integrate sentiment information for enhanced multimodal sarcasm detection.
- **DMSD-CL** [13] aims to alleviate the harmful effects of biased textual factors on robust OOD generalization.
- **DocMSU** [9] introduces a fine-grained sarcasm comprehension method to properly align the pixel level image features.

- **TFCD** [59] proposes a Training-Free Counterfactual Debiasing framework.

The MFND baselines combine images and texts information as follows:

- **EANN** [43] exports event-invariant features for fake news detection through an end-to-end framework.
- **MVAE** [16] uses a bimodal variational autoencoder combined with a binary classifier to perform fake news detection tasks.
- **MCAN** [48] proposes a novel multimodal co-attention network to better fuse text and visual features for fake news detection.
- **CAFE** [6] proposes a cross-modal ambiguity learning problem from the perspective of information theory and performs multimodal fake news detection through ambiguity perception.
- **LIIMR** [38] captures intra-modal relations by extracting fine-grained representations of images and text.
- **SEMI-FND** [37] proposes a novel multimodal stacking integration framework to ensure faster performance with fewer parameters.
- **MFIR** [46] uses multimodal fusion and inconsistent reasoning to discover multimodal inconsistent semantics for explainable fake news detection.
- **CDD** [7] proposes a debiasing framework based on causal intervention and counterfactual reasoning for multimodal fake news detection.
- **CSFND** [31] proposes contextual semantic representation learning to introduce contextual information into the representation learning process.
- **KEVL** [11] integrates information from large-scale open knowledge graphs to augment its ability to discern the veracity of news content.

4.4. Model Comparison

We report the comparative results of various baselines on the MVSA, MSD, and Twitter datasets. We follow prior studies [3, 48, 57] in using standard evaluation metrics. For the MVSA dataset, we assess the model using Accuracy (Acc) and F1 score (F1). On the MSD dataset, we evaluate the model using Accuracy, along with the Binary and Macro Averages of Precision (P), Recall (R), and F1 score. For the MFDN dataset, the model is evaluated based on Accuracy, Precision, Recall, and F1 score.

As illustrated in Table 1, Table 2, and Table 3, the following conclusions can be drawn: From the perspective of task characteristics, our proposed MDPF surpasses all existing baselines across all evaluation metrics on the MVSA, MSD and Twitter datasets. These outcomes suggest that framing the MSA, MSD, and MFND tasks as a sequence that integrates semantics and sentiment leads to more effective multimodal classification. The significance of our results is confirmed through rigorous testing, with $p < 0.05$ indicating statistical significance. From the perspective of the technical characteristics, **1)** MDPF using the origin BERT and ViT outperforms the baselines,

Table 1: Performance of MDPF compared to baselines on MVSA-single and MVSA-multiple with the evaluation metrics Accuracy and F1-score. $B\&V$ denotes the model using origin BERT and ViT, and $CLIP$ denotes the BERT and ViT based on CLIP model. Results with $\star$ denote the significance tests of MDPF over the baseline models at $p < 0.05$.

Models	**MVSA-Single**		**MVSA-Multiple**	
	Accuracy	F1-score	Accuracy	F1-score
CoMN [49]	70.5	70.0	68.9	68.8
FENet [14]	74.2	74.0	71.4	71.2
CM-BERT [52]	71.3	72.7	70.0	73.6
MAG [34]	77.8	76.2	75.1	74.3
MVAN [53]	73.1	72.3	72.4	72.3
ITIN [57]	75.2	75.0	73.5	73.5
DMLANet [51]	79.4	79.5	77.8	75.2
MDPF + B&V*	80.1	80.1	78.4	78.2
MDPF + CLIP*	**80.6**	**80.3**	**79.2**	**78.5**

indicating the effectiveness of the proposed method. Meanwhile, the performance with CLIP initialization surpasses that of using original BERT and ViT approaches, indicating that aligned multimodal representations are more conducive to enhancing the quality of feature fusion. **2)** MDPF surpasses MSA baselines that model multimodal congruity via sophisticated attention mechanisms, such as FENet and ITIN, demonstrating that the dynamic connective strategy within semantic joint perception module is capable of adaptively matching image-text pairs via a routing attention mechanism, thereby effectively capturing representations of multimodal congruity. **3)** MDPF also

Table 2: Performance of MDPF compared to state-of-the-art baselines on MSD with the evaluation metrics Accuracy, Precision, Recall and F1-score. The other settings are consistent with MVSA. Results with $\star$ denote the significance tests of MDPF over the baseline models at $p < 0.05$.

Model	Acc	Binary-Average			Macro-Average		
		P↑	R↑	F1↑	P↑	R ↑	F1↑
HFM[3]	86.6	83.8	84.1	84.0	86.2	86.2	86.2
Res-BERT[30]	84.8	77.8	84.1	80.8	78.8	84.4	81.5
Att-BERT[30]	86.0	78.6	83.3	80.9	80.8	85.0	82.9
InCrossMGs[21]	86.1	81.3	84,3	82.8	85.3	85.8	85.6
CMGCN[22]	87.5	83.6	84.6	84.1	87.0	86.9	87.0
HKEmodel[23]	87.3	81.8	86.4	84.0	-	-	-
MILNet[32]	89.5	85.1	89.1	87.1	88.8	89.4	89.1
DIP[45]	89.5	87.7	86.5	87.1	88.4	89.1	89.0
KnowleNet[54]	88.8	88.5	84.1	86.3	88.8	88.2	88.5
DynRT-Net [39]	88.9	88.6	87.9	88.2	-	-	-
SAHFN [24]	87.2	82.7	87.3	84.9	86.7	87.2	86.9
DMSD-CL [13]	88.9	84.8	87.9	86.3	88.3	88.7	88.5
DocMSU [9]	89.7	86.3	88.4	87.3	88.8	89.2	89.0
TFCD [59]	89.5	84.8	89.4	88.1	-	-	-
MDPF + B&V*	90.2	89.2	88.3	88.7	89.7	89.3	89.5
MDPF + CLIP*	**90.8**	**89.5**	**88.9**	**89.2**	**90.1**	**89.4**	**89.7**

demonstrates optimal performance on the MSD task. Different from MSA task, MSD task aims to identify incongruity between text and image. Therefore, compared to baselines utilizing advanced pre-trained models and complex fusion mechanisms, such as HKEmodel and DynRT-Net, MDPF employs a soft probability based on gaussian distribution to model incongruity. The strategy can avoid the errors and biases that fixed thresholds may introduce. **4)** MFND task is similar to the MSD task and also achieves the best performance. Compared to pre-trained stacking and module integration models such as SEMI-FND, MDPF utilizes cross-modal GCN to preserve the structured information and share information between modalities, introducing sentiment knowledge to refine the edge weights of the graph to better capture cross-modal sentiment interaction. **5)** Compared to baselines that utilize traditional attention-aligned methods during the encoding stage, MDPF demonstrates excellent performance. We consider that MDPF introduces the external knowledge to augment the multimodal semantics and

Table 3: Performance of MDPF compared to state-of-the-art baselines on Twitter with the evaluation metrics Accuracy, Fake News F1-score and Real News F1-score. The other settings are consistent with MVSA. Results with $\star$ denote the significance tests of MDPF over the baseline models at $p < 0.05$.

Model	Acc	Fake News			Real News		
		P↑	R↑	F1↑	P↑	R ↑	F1↑
EANN [43]	64.8	81.0	49.8	61.7	58.4	75.9	66.0
MAVE [16]	74.5	80.1	71.9	75.8	68.9	77.7	73.0
MCAN [48]	79.6	78.5	89.1	83.5	81.9	66.9	73.6
CAFE [6]	80.6	80.7	79.9	80.3	80.5	81.3	80.9
LIIMR [38]	83.1	83.6	83.2	83.0	84.0	84.3	84.1
MFIR[46]	85.8	85.0	84.8	84.9	86.7	87.1	86.9
CDD [7]	87.4	82.0	79.2	80.6	89.9	91.4	90.6
CSFND [31]	83.3	89.9	79.9	84.6	76.3	87.8	81.7
KEVL [11]	826	79.9	81.6	80.7	84.9	83.4	84.2
MDPF + B&V*	88.5	87.6	85.7	86.6	90.1	90.5	90.3
MDPF + CLIP*	**89.4**	**90.2**	**85.5**	**87.8**	**90.3**	**91.6**	**90.9**

comprehensively understand the underlying semantics and interaction via contrastive learning.

4.5. Ablation Study

To evaluate the effectiveness of the main components, we conduct an ablation study to remove each one from proposed MDPF for comparison, as shown in Table 4.

Firstly, we observe that the performance drops significantly across all datasets when remove the MCE module (i.e., (a)), this indicates that obtaining high-quality representations is crucial for enhancing model performance. To further explore this phenomenon, we compared the effects of solely removing the BLK (i.e., (b)) and solely removing the CCL (i.e., (c)), and found the decline of the BLK is significantly greater than CCL. This indicates that CCL relies on BLK. BLK supplements multimodal information in the original modal space by generating image descriptions, laying the groundwork for performing a comprehensive understanding of the underlying semantics and interactions via CCL.

Secondly, the setting of (d) obtains worse results than the original model (i.e., (i)). This reason is that SJP can capture shared and private features to model semantic congruity and incongruity via dynamic connective mechanism and g. Meanwhile, we observe that the setting of (e) experienced a greater decline in performance on the MVSA dataset compared to (f), while the opposite is true for the MSD and Twitter datasets. We consider this to be caused by the characteristics of the tasks, with MVSA focusing more on congruity in semantic modeling, while MSD and Twitter pay more attention to incongruity between text and images. Therefore, when the DC mechanism is removed, the model fails to capture semantic congruity, resulting in a significant decrease in its ability to classify multimodal sentiment. Conversely, when the GD is removed, the model loses its ability to model incongruity between text and images, leading to a decline in its capacity for detecting sarcasm and fake news.

Finally, the performance decline is also significant after removing the SIP module (i.e., (g)). Notably, in the setting (g), the performance decrease on the Twitter dataset is not pronounced. We hypothesize two reasons for this: first, multimodal fake news detection is more concerned with semantic incongruity between text and images,

Table 4: Ablation study of the proposed MDPF on different components. BLK and CCL represent the bootstrapping language-image knowledge and cross-modal contrastive learning from MCE. DC and GD represent the dynamic connective mechanism and gaussian distribution from SJP. ✗ indicates the remove strategy, and ✓ indicates consistency with other settings.

Setings	Components						MVSA-single		MVSA-multiple		MSD		Twitter		
	MCE		SJP		SIP	CFM	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	Fake F1	Real F1
	BLK	CCL	DC	GD											
(a)	✗	✗	✓	✓	✓	✓	75.5	74.7	73.6	73.1	85.8	84.5	84.4	83.4	86.3
(b)	✗	✓	✓	✓	✓	✓	77.0	75.6	75.3	74.7	87.5	85.9	85.2	83.9	87.2
(c)	✓	✗	✓	✓	✓	✓	78.2	76.6	76.8	76.4	88.7	86.9	86.4	84.5	87.6
(d)	✓	✓	✗	✗	✓	✓	76.3	74.9	75.2	75.0	86.1	83.8	86.6	85.3	88.5
(e)	✓	✓	✗	✓	✓	✓	77.8	76.3	76.6	76.1	88.7	86.4	89.0	87.2	89.4
(f)	✓	✓	✓	✗	✓	✓	78.1	76.9	77.3	76.4	87.3	85.0	87.9	86.8	88.9
(g)	✓	✓	✓	✓	✗	✓	77.5	76.2	76.7	75.5	86.6	85.8	87.4	86.5	88.7
(h)	✓	✓	✓	✓	✓	✗	79.4	78.4	77.3	77.2	89.8	88.2	87.9	87.1	89.3
(i)	✓	✓	✓	✓	✓	✓	**80.6**	**80.3**	**79.2**	**78.5**	**90.8**	**89.7**	**89.4**	**87.8**	**90.9**

which is very similar to multimodal sarcasm detection. Second, although sentiment contributes to the detection of fake news information, not all data contain sentiment elements. While cross-modal graph convolutional networks can preserve the structured information of image and text data and share information between modalities, thereby better capturing cross-modal emotional relationships, this may introduce noise when sentiment information is not sensitive. Additionally, the setting (h) involves removing the combination fusion strategy and instead using a direct concatenation approach for multimodal fusion. It is clear that using a combination fusion strategy significantly outperforms the concatenation method.

4.6. Parameter analysis

Since dynamic connection and graph convolution components are related to the construction of semantic intensified and sentiment interactive information, we conduct a parameter analysis to explore the impact of different layers of DC and GCN on the accuracy in three tasks, as shown in Figure 3.

First, as the number of DC layers increases, we observe that the accuracy of MDPF on the three tasks steadily improves and reaches a peak at the 5th layer. This shows that the increase in the number of connection layers enables MDPF to continuously learn the congruity representation of images and texts via a multi-head dynamic routing mechanism to capture the most appropriate image-text pairs. However, as the number of

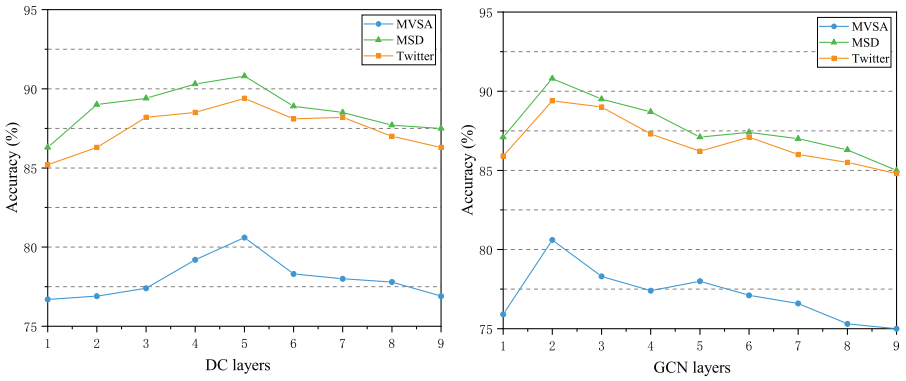

Figure 3: Parameter analysis on the accuracy in MVSA, MSD and MFND tasks.

layers further increases, the dynamic connection components will gradually introduce cross-modal relations that are irrelevant to the current image-text pair, resulting in performance degradation.

Secondly, different from the number of DC layers, the model performance is optimal when the number of GCN layers is set to 2. This shows that the GCN has a strong ability to aggregate information. However, as the number of layers increases, the performance of the model gradually decreases. This phenomenon occurs because the GCNk has an over-smoothing issue, and the increase in the number of layers makes the features of all nodes more similar.

4.7. Case Study

We randomly select three cases from the three datasets to explain how bootstrapping language-image knowledge affects the proposed MDPF model, as shown in Figure 4. First, in the case of MVSA, we observe that the phrase *Day 51 of 90 paintings in 90 Days of Stratford!* highlights *paintings* and *Stratford*. However, neither these two words nor the entire sentence exhibit a clear sentiment attitude. The overall sentiment polarity would overly rely on the image encoding of low-level semantic features, leading to erroneous judgments by the baseline model. When we employ the BLK module to

Datasets	**MVSA**	**MSD**	**Twitter**
Image			
Text	Day 51 of 90 paintings in 90 Days of Stratford!	Gotta loves the warm and dry conditions!	Fantastic! and plenty of cloud for all of us down on earth.
BIK	The painting of a flower garden with a gazebo.	Church with a steeple in the snow and a street covered in snow.	A solar eclipse seen from space.
Baselines	Neutral (False)	Non-Sarcasm (False)	Real (False)
MDPF	Positive (True)	Sarcasm (True)	Fake (True)

Figure 4: Case analysis on MVSA, MSD and MFND datasets.

generate image descriptive knowledge *The painting of a flower garden with a gazebo*, the terms *flower garden* and *gazebo* exhibit a positive sentiment tendency to some extent. MDPF makes accurate decisions with the aid of bright colors in the combined visual features. Then, case MSD presents a sarcasm example. The sentence *Gotta love the warm and dry conditions* expresses a positive sentiment via words *love*, *warm* and *dry* while the overall white background of the image does not provide a good reference for the incongruity between text and image. The image description of *Church with a steeple in the snow and a street covered in snow* contrasts the *snow* with *warm*, aiding the model in accurately identifying the sarcasm incongruity. The case of Twitter is similar to the case of MSD. The BLK-generated *solar eclipse* and *space* create a clear contrast with text words *Fantastic*, *cloud*, and *earth*, helping the model to overcome the deception of the image object *cloud* on the textual semantics of *cloud*. This further demonstrates that the BLK can assist MDFP in identifying fake news incongruity.

4.8. Visualization

To further discuss the effectiveness of our method, we provide a visualization to discuss semantic joint perception and cross-modal sentiment interactive graphs on how to assist the dual perception module in capturing semantic and sentiment information. Firstly, Figure 5 shows the impact of the semantic joint perception module. The dynamic connection mechanism allows the model to focus on cross-modal regions related to the text by modeling shared semantics, such as exhibiting high attention to image objects *Strawberry*, *Nutella*, and the word *happiness*. Meanwhile, the gaussian-weighted distribution pays more attention to private semantics to achieve semantic inconsistency modeling, as manifested by objects like *cup* in the image being identified and distinguished.

Secondly, Figure 6 presents an example of constructing a cross-modal graph in the sentiment interaction perception module. We can observe that the positive sentiment expressed by *positively* contrasts sharply with the gloomy colors in the hurricane and dark clouds. The cross-modal sentiment interaction is represented by edges with greater weights to represent highly relevant sentiment clues, indicating the effectiveness of cross-modal graph convolution in capturing multimodal sentiment interactions.

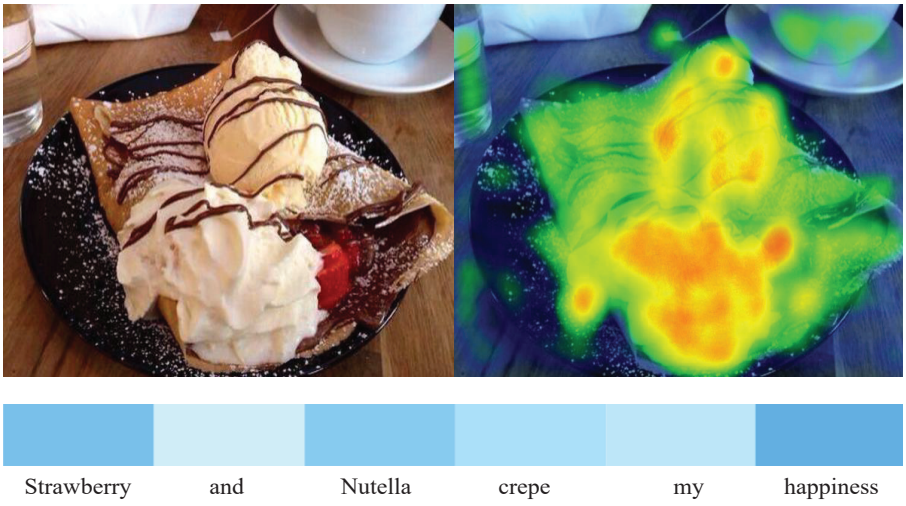

Figure 5: Attention visualization of the semantic joint perception mechanism.

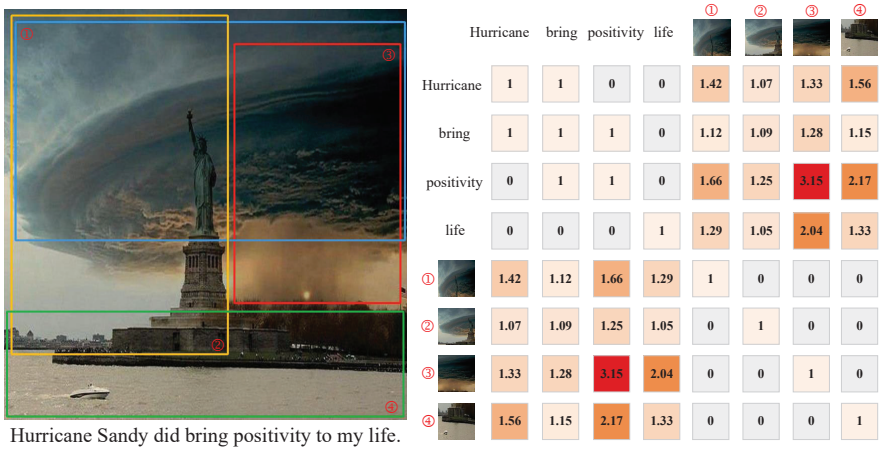

Figure 6: Sentiment interactive matrix of the cross-modal graph which presents the key word.

5. Conclusion

In this paper, we introduce MDPF, a pioneering unified multimodal fusion framework designed for a broader spectrum of multimodal affective analysis, encompassing sentiment analysis, sarcasm detection, and fake news detection. This framework innovatively constructs a fusion sequence of semantics and sentiment to enhance the

performance of multimodal affective classification. Initially, we employ a bootstrapping language-image knowledge base and cross-modal contrastive learning to enrich the original modality space and align cross-modal information. Subsequently, we establish a dynamic connective mechanism that adaptively matches image-text pairs to capture shared semantics, while also employing a gaussian-weighted approach to emphasize private semantics. Concurrently, sentiment knowledge is integrated to refine the edge weights of the cross-modal graph, facilitating the capture of sentiment interactions. Ultimately, we devise a cross-modal combination fusion strategy to merge semantically enriched and sentimentally interactive sequences for comprehensive multimodal classification. The efficacy of the proposed MDPF framework is validated on three publicly available datasets, with empirical results underscoring the superiority of our model.

Acknowledgements

This work was supported by 2022 Yulin Science and Technology Plan Project under Grant No.CXY-2022-177, the Alan Turning Institute Grant: Improving multimodality misinformation detection with affective analysis.

References

[1] Bastick, Z., 2021. Would you notice if fake news changed your behavior? an experiment on the unconscious effects of disinformation. Computers in human behavior 116, 106633.

[2] Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y., 2018. Detection and visualization of misleading content on twitter. International Journal of Multimedia Information Retrieval 7, 71–86.

[3] Cai, Y., Cai, H., Wan, X., 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model, in: Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 2506–2515.

[4] Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A., 2017. Affective computing and sentiment analysis. A practical guide to sentiment analysis , 1–10.

[5] Cambria, E., Li, Y., Xing, F.Z., Poria, S., Kwok, K., 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis, in: Proceedings of the 29th ACM international conference on information & knowledge management, pp. 105–114.

[6] Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., Shang, L., 2022. Cross-modal ambiguity learning for multimodal fake news detection, in: Proceedings of the ACM web conference 2022, pp. 2897–2905.

[7] Chen, Z., Hu, L., Li, W., Shao, Y., Nie, L., 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 627–638.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations.

[9] Du, H., Nan, G., Zhang, S., Xie, B., Xu, J., Fan, H., Cui, Q., Tao, X., Jiang, X., 2024. Docmsu: A comprehensive benchmark for document-level multimodal sarcasm understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 17933–17941.

[10] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A., 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion 91, 424–444.

[11] Gao, X., Wang, X., Chen, Z., Zhou, W., Hoi, S.C., 2024. Knowledge enhanced vision and language model for multi-modal fake news detection. IEEE Transactions on Multimedia .

[12] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for

unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

[13] Jia, M., Xie, C., Jing, L., 2024. Debiasing multimodal sarcasm detection with contrastive learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 18354–18362.

[14] Jiang, T., Wang, J., Liu, Z., Ling, Y., 2020. Fusion-extraction network for multimodal sentiment analysis, in: Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24, Springer. pp. 785–797.

[15] Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J., 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM international conference on Multimedia, pp. 795–816.

[16] Khattar, D., Goud, J.S., Gupta, M., Varma, V., 2019. Mvae: Multimodal variational autoencoder for fake news detection, in: The world wide web conference, pp. 2915–2921.

[17] Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

[18] Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR. pp. 19730–19742.

[19] Li, S., Ham, J., Eastin, M.S., 2024. Social media users' affective, attitudinal, and behavioral responses to virtual human emotions. Telematics and Informatics 87, 102084.

[20] Liang, B., Gui, L., He, Y., Cambria, E., Xu, R., 2024. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. IEEE Transactions on Affective Computing .

[21] Liang, B., Lou, C., Li, X., Gui, L., Yang, M., Xu, R., 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: Proceedings of the 29th ACM international conference on multimedia, pp. 4707–4715.

[22] Liang, B., Lou, C., Li, X., Yang, M., Gui, L., He, Y., Pei, W., Xu, R., 2022. 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), Association for Computational Linguistics. pp. 1767–1777.

[23] Liu, H., Wang, W., Li, H., 2022. 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, pp. 4995–5006.

[24] Liu, H., Wei, R., Tu, G., Lin, J., Liu, C., Jiang, D., 2024. Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. Information Fusion 108, 102353.

[25] Lu, Q., Long, Y., Sun, X., Feng, J., Zhang, H., 2024a. Fact-sentiment incongruity combination network for multimodal sarcasm detection. Information Fusion 104, 102203.

[26] Lu, Q., Sun, X., Gao, Z., Long, Y., Feng, J., Zhang, H., 2024b. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Information Processing & Management 61, 103538.

[27] Luo, J., Borth, D., You, Q., 2017. Social multimedia sentiment analysis, in: Proceedings of the 25th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 1953–1954. doi:10.1145/3123266.3130143.

[28] Niu, T., Zhu, S., Pang, L., El Saddik, A., 2016. Sentiment analysis on multi-view social data, in: MultiMedia Modeling: 22nd International Conference, MMM

2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22, Springer. pp. 15–27.

[29] Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

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

[31] Peng, L., Jian, S., Kan, Z., Qiao, L., Li, D., 2024. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. Information Processing & Management 61, 103564.

[32] Qiao, Y., Jing, L., Song, X., Chen, X., Zhu, L., Nie, L., 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9507–9515.

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

[34] Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E., 2020. Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, NIH Public Access. p. 2359.

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

[36] Schifanella, R., de Juan, P., Tetreault, J., Cao, L., 2016b. Detecting sarcasm in multimodal social platforms, in: Proceedings of the 24th ACM International Conference on Multimedia, Association for Computing Machinery, New York, NY, USA. p. 1136–1145. doi:10.1145/2964284.2964321.

[37] Singh, P., Srivastava, R., Rana, K., Kumar, V., 2023. Semi-fnd: Stacked ensemble based multimodal inferencing framework for faster fake news detection. Expert systems with applications 215, 119302.

[38] Singhal, S., Pandey, T., Mrig, S., Shah, R.R., Kumaraguru, P., 2022. Leveraging intra and inter modality relationship for multimodal fake news detection, in: Companion Proceedings of the Web Conference 2022, pp. 726–734.

[39] Tian, Y., Xu, N., Zhang, R., Mao, W., 2023. Dynamic routing transformer network for multimodal sarcasm detection, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2468–2480.

[40] Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems .

[41] Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. science 359, 1146–1151.

[42] Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., Wang, S., 2023. Cross-modal contrastive learning for multimodal fake news detection, in: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5696–5704.

[43] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J., 2018. Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp. 849–857.

[44] Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., Li, D., 2022. Cross-modal knowledge distillation in multi-modal fake news detection, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4733–4737.

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

[46] Wu, L., Long, Y., Gao, C., Wang, Z., Zhang, Y., 2023. Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. Information Fusion 100, 101944.

[47] Wu, Y., Lin, Z., Zhao, Y., Qin, B., Zhu, L.N., 2021a. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4730–4738.

[48] Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z., 2021b. Multimodal fusion with co-attention networks for fake news detection, in: Findings of the association for computational linguistics: ACL-IJCNLP 2021, pp. 2560–2569.

[49] Xu, N., Mao, W., Chen, G., 2018. A co-memory network for multimodal sentiment analysis, in: The 41st international ACM SIGIR conference on research & development in information retrieval, pp. 929–932.

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

[51] Yadav, A., Vishwakarma, D.K., 2023. A deep multi-level attentive network for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications and Applications 19, 1–19.

[52] Yang, K., Xu, H., Gao, K., 2020a. Cm-bert: Cross-modal bert for text-audio sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, pp. 521–528.

[53] Yang, X., Feng, S., Wang, D., Zhang, Y., 2020b. Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia 23, 4014–4026.

[54] Yue, T., Mao, R., Wang, H., Hu, Z., Cambria, E., 2023. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. Information Fusion 100, 101921.

[55] Zhang, H., Fang, Q., Qian, S., Xu, C., 2019. Multi-modal knowledge-aware event memory network for social media rumor detection, in: Proceedings of the 27th ACM international conference on multimedia, pp. 1942–1951.

[56] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al., 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 .

[57] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J., 2022. Multimodal sentiment analysis with image-text interaction network. IEEE transactions on multimedia .

[58] Zhu, Z., Zhang, D., Li, L., Li, K., Qi, J., Wang, W., Zhang, G., Liu, P., 2023. Knowledge-guided multi-granularity gcn for absa. Information Processing & Management 60, 103223.

[59] Zhu, Z., Zhuang, X., Zhang, Y., Xu, D., Hu, G., Wu, X., Zheng, Y., 2024. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing, in: Proc. of IJCAI.

[60] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R., 2018. Detection and resolution of rumours in social media: A survey. Acm Computing Surveys (Csur) 51, 1–36.