Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061401

ARTICLE

PKME-MLM: A Novel Multimodal Large Model for Sarcasm Detection

Jian Luo[1], **Yaling Li**[1], **Xueyu Li**[1] and **Xuliang Hu**[2,*]

[1]College of Information Science and Engineering, Hunan Normal University, Changsha, 410000, China
[2]Institute of Interdisciplinary Studies, Hunan Normal University, Changsha, 410000, China

Corresponding Author: Xuliang Hu. Email: huxuliang@hunnu.edu.cn

Received: 23 November 2024; Accepted: 02 January 2025; Published: 26 March 2025

ABSTRACT: Sarcasm detection in Natural Language Processing (NLP) has become increasingly important, particularly with the rise of social media and non-textual emotional expressions, such as images. Existing methods often rely on separate image and text modalities, which may not fully utilize the information available from both sources. To address this limitation, we propose a novel multimodal large model, i.e., the PKME-MLM (Prior Knowledge and Multi-label Emotion analysis based Multimodal Large Model for sarcasm detection). The PKME-MLM aims to enhance sarcasm detection by integrating prior knowledge to extract useful textual information from images, which is then combined with text data for deeper analysis. This method improves the integration of image and text data, addressing the limitation of previous models that process these modalities separately. Additionally, we incorporate multi-label sentiment analysis, refining sentiment labels to improve sarcasm recognition accuracy. This design overcomes the limitations of prior models that treated sentiment classification as a single-label problem, thereby improving sarcasm recognition by distinguishing subtle emotional cues from the text. Experimental results demonstrate that our approach achieves significant performance improvements in multimodal sarcasm detection tasks, with an accuracy (Acc.) of 94.35%, and Macro-Average Precision and Recall reaching 93.92% and 94.21%, respectively. These results highlight the potential of multimodal models in improving sarcasm detection and suggest that further integration of modalities could advance future research. This work also paves the way for incorporating multimodal sentiment analysis into sarcasm detection.

KEYWORDS: Sarcasm detection; multimodal large model; prior knowledge; multi-label fusion

1 Introduction

Sarcasm is a form of emotional expression that highlights the disparity between a person's true intentions and the content they explicitly present [1], indicating that the true attitude is contrary to the literal meaning [2]. With the rapid growth of social media and online communication, people are increasingly expressing emotions through various forms, such as text and images, which provides a rich data resource for multimodal sarcasm detection. Sarcasm Detection (SD) has received growing attention due to its wide range of applications [3]. Multimodal sarcasm detection leverages these different forms of data to more accurately identify and understand human emotions and attitudes. It goes beyond traditional text-based sarcasm detection [4] by incorporating the analysis of modalities such as images, enabling sarcasm detection to understand human complex emotional expressions more comprehensively. This approach of integrating multiple data sources opens new possibilities for applications in fields such as NLP [5], intelligent interaction [6], and social media monitoring [7], improving the accuracy and scope of sarcasm detection in sentiment analysis.

Multimodal sarcasm detection (MSD) faces two main challenges. First, the primary challenge lies in learning cross-modal fusion representations, which often focus solely on exploring interactions between modalities, thereby neglecting the disharmony among them. To address this, Wu et al. [8] proposed an Incongruity-aware Weighted Attention Network (IWAN), which uses a scoring mechanism to detect sarcasm by focusing on word-level incongruence between modalities, thus assigning greater weight to inconsistent words. Second, another challenge is that most existing work in MSD relies on limited superficial information, overlooking the integration of contextual knowledge, which limits their ability to achieve better multimodal sarcasm detection. Therefore, Amir et al. [7] proposed a deep neural network for automatic sarcasm detection, which leverages contextual features rather than lexical and syntactic cues present in the discourse, thereby enabling intelligent learning and highlighting of significant features of the relevant emotional context in the text. Although these two challenges have been partially addressed, research has found that existing MSD still fails to fully exploit the information present in images and text, leading to a lack of deep integration and extraction of multimodal sarcasm information. Moreover, the simplification of sentiment labels makes precise sarcasm detection challenging.

Sarcasm has been a longstanding topic in various fields such as psychology [9], political science [10], and sociology [11]. To address these issues, we focus on the insufficient exploration and fusion of information in MSD. We propose a model named Text-Image Fusion Summary (TIFS). The goal of this model is to fully exploit all the information present in both images and text, ensuring that most of the information is effectively utilized. Specifically, the TIFS model first divides images into those with text and those without text, based on prior knowledge, and then extracts all visible text information from images containing text. This text information may directly express emotions, suggest sarcasm, or include other valuable information related to the image content. The TIFS model significantly expands the sources of textual data, allowing potential information in images to be effectively captured and utilized. After extracting text from the images, the concatenation module of the TIFS model merges the text extracted from the images with the original text to form a comprehensive text sequence. During the concatenation process, the TIFS model ensures that the newly generated text sequence retains the contextual information of the original text while enhancing the richness of semantic expression. This approach not only integrates content from multiple information sources but also compensates for the limitations of single-text or image information, enabling the full utilization of potential sarcastic information embedded in images and text. This comprehensive text sequence not only enriches the diversity of input information but also improves the model's ability to capture sarcastic information. In the subsequent sarcasm detection process, the TIFS model performs an in-depth analysis of the comprehensive text sequence, enabling a more thorough understanding of the text content. Specifically, the TIFS model ensures that all information extracted from images and text is effectively utilized through multi-level semantic fusion and sentiment analysis.

To address the issue of simplified sentiment labels, this paper introduces a Twitter-specific model [12], which is particularly applied to sarcasm detection tasks. Traditional sentiment analysis and sarcasm detection methods are usually limited to identifying a single sentiment label, which proves inadequate when dealing with complex emotional expressions. Especially on social media platforms like Twitter, where users' tweets are often short yet may contain multiple emotions simultaneously, such as expressing anger while also being sarcastic, this complex emotional expression requires a more flexible and precise analysis tool. This model utilizes a multi-label classification approach, enabling a post to be assigned multiple sentiment labels, thereby providing a more accurate representation of the complex emotions expressed in the text. This model adopts a multi-label classification approach, allowing a tweet to be associated with multiple sentiment labels, thereby more accurately reflecting the complex emotions within the tweet. Specifically, the model effectively addresses the problem of simplified sentiment labels through an innovative multi-label classification approach, providing a new perspective and technical means for multimodal sarcasm detection. This model not only improves the accuracy and robustness of sarcasm detection but also offers new research directions and technical methods for sentiment analysis and information fusion in multimodal NLP tasks.

Therefore, this paper proposes an innovative model named Text-Image Fusion Summary (TIFS) by leveraging prior knowledge and introduces a Twitter-specific model to address the issues of underutilization of image-text information and simplified sentiment labels in current multimodal sarcasm detection. Overall, by combining the TIFS model and the Twitter-specific model to perform sarcasm recognition on images and text, as shown in Fig. 1 (Figure (a) shows a "non-sarcastic" multimodal instance, while Figure (b) shows a "sarcastic" instance. These two examples use different image and text information to demonstrate sarcastic and non-sarcastic scenarios, respectively), more precise sarcasm detection results are achieved. This also provides new research directions and technical methods for sentiment analysis and information fusion in multimodal NLP tasks, promoting further development in this field.

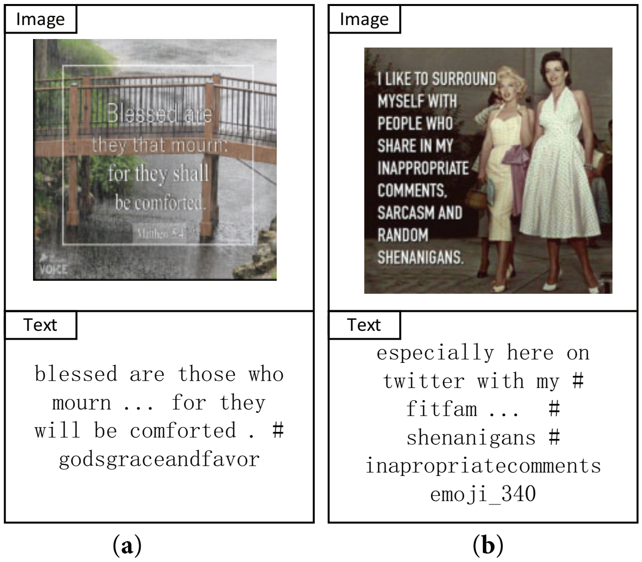


Image
Blessed are they that mourn: for they shall be comforted.
Matthew 5:4
VOICE
Text
blessed are those who mourn ... for they will be comforted . # godsgraceandfavor
(a)
Image
I LIKE TO SURROUND MYSELF WITH PEOPLE WHO SHARE IN MY INAPPROPRIATE COMMENTS, SARCASM AND RANDOM SHENANIGANS.
Text
especially here on twitter with my # fitfam ... # shenanigans # inapropriatecomments emoji_340
(b)

Figure 1: Two examples for sarcasm recognition

Our contributions can be summarized as:

(1) We propose a Text-Image Fusion Summary (TIFS) model, which utilizes prior knowledge to extract textual information from images and concatenates it with the original text, resulting in a more comprehensive summarized text sequence. This method improves the integration of image and text data, addressing the limitation of previous models that process these modalities separately.

(2) We introduced a Twitter-specific model, which has been fine-tuned on a large-scale Twitter dataset and employs a multi-label classification approach, allowing a tweet to have multiple sentiment labels simultaneously. This design enhances sentiment accuracy, overcoming the issue in prior models that treat sentiment classification as a single-label problem, making sarcasm recognition more precise.

(3) This paper systematically verifies the effectiveness and superiority of the proposed method through extensive experimental comparisons and ablation studies. A series of experiments on a public multimodal sarcasm detection dataset demonstrate that the proposed method achieves high performance.

The remainder of this paper is divided into four sections. Section 2 briefly reviews related work, including sarcasm detection and sentiment analysis. Section 3 introduces the PKME-MLM model, including Text-Image Fusion Summary Modeling, Semantic Intensified Distribution Modeling, and Siamese Sentiment Contrastive Learning. Section 4 analyzes the experimental results. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Sarcasm Detection

With the development of multimedia technology and the popularization of social media, sarcasm (SD) has become an important way for users to express their true attitudes and opinions. In early research, scholars found that sarcasm was primarily conveyed through text. This linguistic phenomenon relies on the clever use of words, employing rhetorical devices such as metaphors, irony, and exaggeration, which contrast the literal meaning with the actual intent [13], thereby achieving a sarcastic effect.

With the advancement of NLP technology, researchers such as Joshi et al. [14] began exploring how to automatically detect sarcasm in text. Early methods primarily relied on handcrafted feature extraction, utilizing lexical, syntactic, and semantic features, and employing traditional machine learning algorithms such as Support Vector Machines (SVM) and Naive Bayes classifiers to identify sarcastic elements in text. For instance, Mukherjee et al. [15] used Naive Bayes and fuzzy clustering for sarcasm detection in microblogs, achieving an accuracy of around 65%, which was considered high for sarcasm detection at that time. However, the shortcomings of this method include its reliance on manually crafted features and limited ability to capture complex patterns in text, which restricted the model's performance, especially with more nuanced sarcasm. Sarsam et al. [16] applied machine learning algorithms like SVM for sarcasm detection on Twitter, which improved the performance of sarcasm detection and proved useful for future researchers or machine learning developers. However, the performance of these methods was often limited by the expertise of the feature designers and the diversity of the datasets.

To overcome these limitations, researchers gradually turned to deep learning techniques, particularly neural network-based methods. In 2018, Porwal et al. [17] used a Recurrent Neural Network (RNN) model for sarcasm detection. The study found that the RNN model could automatically extract features required for machine learning methods. However, this approach has limitations in capturing long-range dependencies and subtle sarcastic expressions, which restricts its performance. Subsequently, in 2020, Salim et al. [18] employed a deep LSTM-RNN with word embedding capabilities for sarcasm detection on Twitter. The advantage of this approach was the ability to capture long-term dependencies and the use of word embeddings to better understand semantic meanings. This made sarcasm detection on Twitter more effective, achieving improved performance. Despite its success, this method still faced challenges in handling noisy or informal language commonly found in tweets, limiting its robustness. Kumar et al. [19] introduced a multi-head attention BiLSTM (MHA-BiLSTM) network to enhance the performance of BiLSTM, outperforming feature-rich SVM models. In recent years, although significant progress has been made in text-based sarcasm detection, traditional sarcasm detection methods (Joshi et al. [20]) often considered only textual information while ignoring potential emotional information from other modalities such as images, which limited the accuracy and comprehensiveness of detection.

To further improve the accuracy of sarcasm detection, researchers have gradually shifted towards multimodal sarcasm detection (MSD). MSD is a technique that utilizes information from multiple modalities, such as text and images, for emotion recognition, aiming to achieve more comprehensive and precise sarcasm detection on a given multimodal dataset. In 2022, Ding et al. [21] proposed a multimodal fusion method for sarcasm detection based on late fusion, which improved the performance of sarcasm detection. However, the limitation of this method lies in treating each modality independently, which may overlook the potential interactions between them. In 2023, Yue et al. [22] introduced a new model named KnowleNet, which integrates prior knowledge using the ConceptNet knowledge base and determines image-text relevance through sample-level and word-level cross-modal semantic similarity detection, thereby enhancing the accuracy of sarcasm detection.

Although multimodal techniques have been widely applied in sarcasm detection, the information within different modalities is often not fully explored and utilized, making reliance solely on subsequent multimodal fusion insufficient. Existing methods typically focus on simply fusing information from different modalities while neglecting the potential complex relationships and deep semantics within each modality. This limitation presents opportunities, leaves room for enhancing improvement in the accuracy and robustness of sarcasm detection.

To address this issue, we introduce prior knowledge and propose a novel Text-Image Fusion Summary (TIFS) method. This method not only focuses on the shallow superficial fusion of multimodal information but also extracts and integrates textual information from images, concatenating it with the original text for deep fusion, thereby generating a more comprehensive and precise fused text. The core of the TIFS method is to fully exploit the potential information within each modality and, guided by prior knowledge, enable more efficient and effective information interaction between different modalities.

2.2 Sentiment Analysis

Sentiment analysis is closely related to sarcasm detection [23]. As a prominent and widely applied research area, sentiment analysis holds an important position in NLP. By analyzing the sentiment tendencies within text, sentiment analysis can reveal users' attitudes and opinions [24], which is crucial for understanding sarcastic expressions.

Early research on sentiment analysis mainly relied on manually constructed sentiment lexicons, such as SentiWordNet [25], which associated words with specific sentiment polarities (e.g., positive, negative, or neutral) to achieve preliminary classification of text sentiment. With the rapid development of machine learning techniques, researchers began to adopt machine learning approaches to solve sentiment analysis problems. In 2014, Zainuddin et al. [26] used Support Vector Machine (SVM) to train a sentiment classifier and employed N-gram and different weighting schemes to extract classic features, thereby improving classification accuracy. The approach effectively captured sentiment patterns using traditional machine learning techniques. However, its reliance on manually crafted features limits its ability to capture the complexities of language, especially subtle sentiments in larger or more diverse datasets. In the same year, Parmar et al. [27] utilized Random Forest with tuned hyperparameters for sentiment mining. The approach demonstrated that hyperparameter optimization could significantly improve the robustness of sentiment classification. However, this method has limitations in terms of scalability and may face performance issues when applied to large, real-time datasets due to the computational cost of Random Forest models. Later, in 2016, Goel et al. [28] used SentiWordNet and Naive Bayes to improve tweet classification accuracy by providing positivity, negativity, and objectivity scores for words present in tweets. This method used sentiment lexicons to improve classification, but its reliance on static resources limits adaptability to evolving language and domain-specific terms.

In recent years, the rise of deep learning technology has brought new opportunities to sentiment analysis. Liao et al. [29] proposed a method for understanding real-world situations through sentiment analysis of a Twitter database based on deep learning techniques. However, the method's reliance on CNNs might result in high computational costs when processing large volumes of data, which could limit its scalability and efficiency in real-time applications. Li et al. [30] introduced an LSTM-based RNN language model that effectively captures the complete sequence information. However, it may struggle with long texts or noisy data, affecting scalability and robustness in real-time sentiment analysis.

Additionally, with the development of large models, Naseem [31] and others introduced the Transformer architecture into sentiment analysis, further advancing progress in this field. Models such as BERT and RoBERTa, which employ bidirectional encoding of text, can capture richer contextual information. For example, Liu et al. [32] proposed a BERT-based cross-domain aspect-level sentiment analysis algorithm, achieving fine-grained sentiment analysis across domains. Liao et al. [33] introduced a multi-task aspect-category sentiment analysis model based on RoBERTa, which extracts features from text and aspect markers using a deep bidirectional transformer and applies a cross-attention mechanism to guide the model to focus on features most relevant to the given aspect category. These models have achieved remarkable results in various NLP tasks and have become one of the mainstream approaches in the field of sentiment analysis.

Despite the significant progress made by the Transformer architecture and its derived models in sentiment analysis, sarcasm detection, as a complex sentiment analysis task, still faces many challenges. Sarcasm often contains multiple emotions, such as anger, disappointment, and humor, which further complicates the detection process. Traditional sentiment analysis models often struggle with multi-emotion label issues, failing to effectively distinguish and identify the multiple emotions embedded in sarcasm.

To address this issue, we introduced a Twitter-specific model. This model is a deep learning model specifically designed for multi-emotion label problems, capable of simultaneously recognizing and classifying multiple emotions in text. Unlike traditional models, this model not only captures a single emotion in the text but also handles complex multi-emotion expressions, such as mixed emotions in sarcastic texts. This multi-emotion labeling capability allows the Twitter-specific model to perform excellently in sarcasm detection, enabling a more comprehensive recognition and classification of different emotional dimensions in sarcastic texts, thus enhancing the accuracy and robustness of sarcasm detection.

3 Methodology

3.1 Overview

The overall framework structure of PKME-MLM is shown in Fig. 2.

Fig. 2 shows the overall structure of the multimodal sarcasm detection model PKME-MLM, including the TIFS module, SID branch, and SSC branch. The methodology of the multimodal sarcasm detection model PKME-MLM can be summarized as the following steps:

(1) Text-Image Fusion Summary (TIFS) Module: First, the TIFS module extracts text through the Text Extraction Section (TES), which integrates image and text information. The Content Summary Section (CSS) then generates a summarized text sequence that fully incorporates both image and text data.

(2) Feature Extraction: The image and text features are extracted using the Vision Transformer (ViT) for image data and RoBERTa for textual data. These features are then processed and interact through a relation matrix to facilitate cross-modal learning.

(3) SID Branch: The SID branch applies a reweighting strategy and Gaussian distribution modeling to capture sarcastic content from the multimodal features.

(4) SSC Branch: The SSC branch generates multimodal sentiment embeddings through a Siamese network and contrastive learning, which improves sentiment analysis for sarcasm detection.

(5) Fusion and Output: Finally, the features from the SID and SSC branches are concatenated and fused to produce the final output for sarcasm detection.

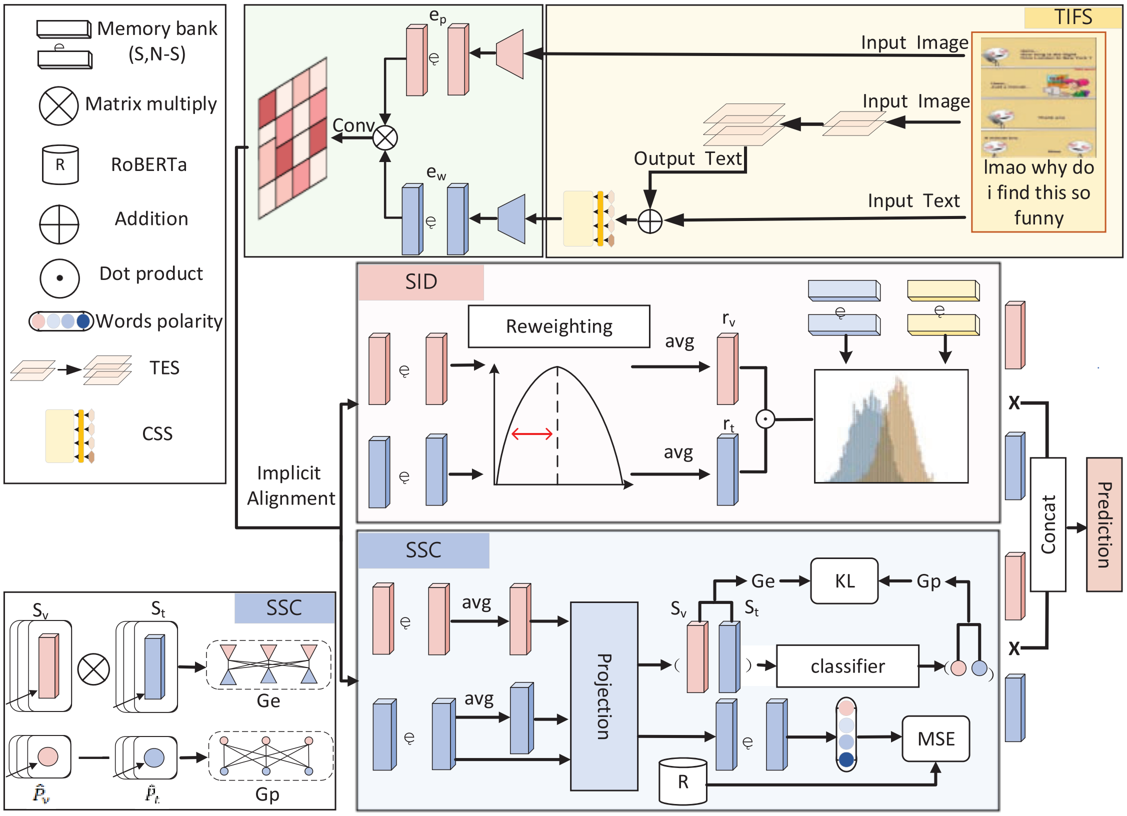


Figure 2: Framework of PKME-MLM

The TIFS module extracts text through the Text Extraction Section (TES) to fully integrate image and text information, and then generates text summaries through the Content Summary Section (CSS). The TIFS module generates a text summary by fully integrating image and text information. Subsequently, image and text features are extracted by ViT (Vision Transformer) and RoBERTa, respectively, and interact through a relation matrix. The SID branch employs a reweighting strategy and Gaussian distribution modeling to capture sarcastic content, while the SSC branch generates multimodal sentiment embeddings through a Siamese network and contrastive learning. Finally, these features are concatenated and fused to produce the output for sarcasm detection. Finally, these features are fused to output the sarcasm detection result.

PKME-MLM is composed of three parts. First, we input the image and text data into the TIFS module, which extracts textual information from the image and performs concatenation and deep fusion with the original text to generate a more comprehensive and precise fused text, ultimately producing a summarized text sequence. Next, we employ a cross-modal implicit alignment strategy to align the summarized text and image and pass the aligned embeddings to the SID and SSC submodules. SID focuses on mining sarcasm information from the factual level, while SSC extracts sarcasm information from the emotional level.

At the factual level, we introduce a channel weighting strategy to obtain embeddings with semantic discrimination, and we use Gaussian distribution to model the correlation uncertainty caused by incongruity. This Gaussian distribution is generated using the latest data stored in a memory bank, which adaptively models the semantic similarity differences between sarcastic and non-sarcastic data.

At the emotional level, we use Siamese layers with shared parameters to learn cross-modal emotional information and construct a polarity value relation graph based on this. To further enhance the extraction ability of emotional information, we introduce a Twitter-specific model that uses a multi-label classification approach, allowing a tweet to have multiple sentiment labels simultaneously. This model captures more precise emotional embeddings through continuous contrastive loss. This multi-level design significantly improves the effectiveness of multimodal sarcasm detection, enabling the model to capture and understand the complex information within sarcastic expressions more comprehensively.

These modules work in collaboration, allowing sarcasm detection to more accurately capture implicit sarcastic information between text and images. Through the comprehensive text sequence generated by the TIFS module, the model can fully leverage cross-modal information, while the SID and SSC sub-modules deeply explore the essence of sarcasm from both factual and emotional perspectives. Such a design not only enhances the overall performance of the model but also improves its ability to detect sarcasm in diverse contexts, ultimately achieving more accurate and comprehensive sarcasm detection.

3.2 Text-Image Fusion Summary Modeling

TIFS consists of two parts: a text extraction section at the head and a content summary section at the tail, as shown in Fig. 3. The text extraction part is constructed using a Transformer architecture, including an image Transformer for extracting visual features and a text Transformer for language modeling. In text-image recognition, we used a standard Transformer encoder-decoder structure.

The specific structure of TIFS is shown in Fig. 3. In the Text Extraction Section, the image is divided into patches and embedded, and then processed through the multi-layer attention mechanism of the encoder and decoder to generate image-text embeddings. The Content Summary Section uses the BERT model to embed and encode the tokenized text. Finally, the features of the image and text are combined to generate the final text summary, achieving the fusion of image and text information.

(1) Text Extraction Section (TES)

The encoder receives the input image $x_{img} \in R^{3\times H_0 \times W_0}$ and resizes it to a fixed size of (H, W). Since the Transformer encoder cannot directly process raw images, the image needs to be transformed into a sequence of input tokens. To achieve this, the encoder divides the resized image into $N = HW/P^2$ square patches of size (P, P). To ensure that the patches fully cover the image region, the width W and height H of the image must be divisible by the patch size p. Then, these patches are flattened into one-dimensional vectors, denoted as $Emb\,(Token_i)$, each token representing the pixel information of an image patch in vector form. These tokens are then projected into a D-dimensional vector space via a linear projection $Proj$, where D represents the hidden dimension size in each layer of the Transformer model. Finally, these vectors become hidden states h_i, as follows:

$$h_i = Proj(Emb\,(Token_i)) \tag{1}$$

Similar to the ViT (Vision Transformer) [34] and DeiT (Data-efficient Image Transformer) [35] models, we retain a special “[CLS]” token in the input sequence. This token plays an important role in image classification tasks as it aggregates the information from all patch embeddings and serves as a global feature representation of the entire image. Additionally, when initializing the encoder using the DeiT pre-trained model, a “distillation token” is introduced into the input sequence to enable the model to learn from a teacher model. These patch embeddings and the two special tokens are assigned learnable one-dimensional position embeddings based on their absolute positions to preserve spatial information.

The task of the decoder is to convert the image information into text. The decoder adopts the original Transformer decoder structure, consisting of a stack of identical layers, similar to the encoder layers, but with one key difference: between the multi-head self-attention and the feed-forward network, the decoder inserts an “encoder-decoder attention” module to assign different attention to different parts of the encoder’s output. In this module, the keys and values of the attention mechanism are derived from the encoder’s output, while the query comes from the decoder’s input.

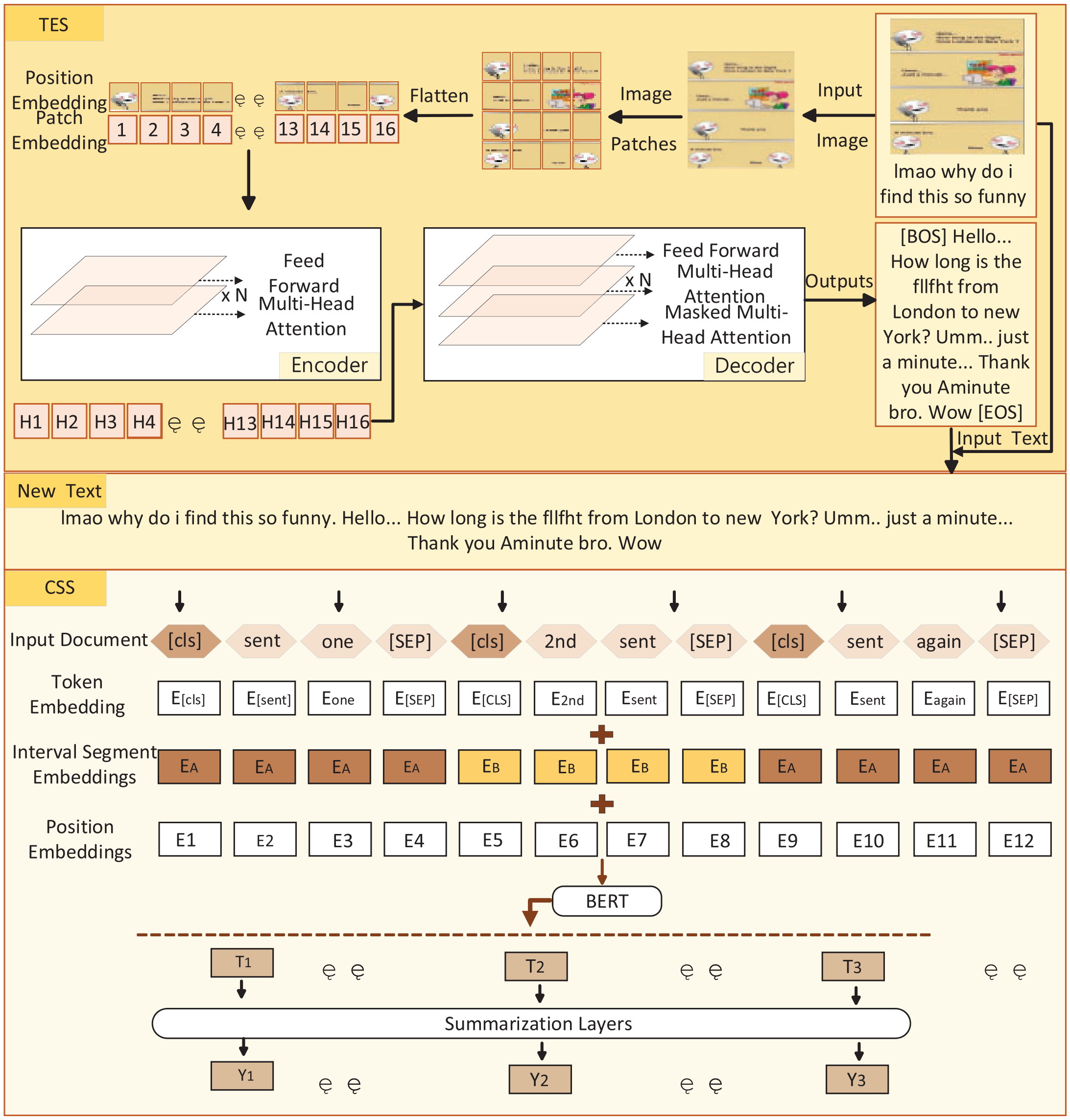

TES
Position Embedding
Patch Embedding
1 2 3 4
13 14 15 16
Flatten
Image Patches
Input Image
lmao why do i find this so funny
[BOS] Hello... How long is the fllfht from London to new York? Umm.. just a minute... Thank you Aminute bro. Wow [EOS]
Input Text
Feed Forward
Multi-Head Attention
x N
Encoder
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
x N
Decoder
Outputs
H1 H2 H3 H4
H13 H14 H15 H16
New Text
lmao why do i find this so funny. Hello... How long is the fllfht from London to new York? Umm.. just a minute... Thank you Aminute bro. Wow
CSS
Input Document
[cls] sent one [SEP] [cls] 2nd sent [SEP] [cls] sent again [SEP]
Token Embedding
E[cls] E[sent] Eone E[SEP] E[CLS] E2nd Esent E[SEP] E[CLS] Esent Eagain E[SEP]
Interval Segment Embeddings
EA EA EA EA EB EB EB EB EA EA EA EA
Position Embeddings
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12
BERT
T1 T2 T3
Summarization Layers
Y1 Y2 Y3

Figure 3: Detailed structure of TIFS

Moreover, the decoder uses attention masking to prevent it from accessing information beyond what is predicted during training. Since the decoder’s output is right-shifted by one position relative to the input, the attention mask needs to ensure that when generating the output at position i, it can only attend to the previous positions (i.e., inputs at positions less than i). The specific formula is as follows:

$$\sigma\left(h_{ij}\right)=\frac{e^{h_{ij}}}{\sum_{k=1}^{V} e^{h_{ik}}} for\ j=1,2,\ldots,V \tag{2}$$

Here, h_{ij} represents the hidden state corresponding to the jth word in the vocabulary, and V is the size of the vocabulary. The hidden states of the decoder are projected from the model dimension to the vocabulary size V through a linear layer, and then the probability of each word is computed using the Softmax function. The final output is obtained through beam search.

The decoder generates the text output required for the OCR task through a series of processing steps, including the encoder-decoder attention mechanism, self-attention masking, linear projection, and Softmax.

(2) Content Summary Section (CSS)

The text extracted from the head image is concatenated with the original text to form a new text. We used the BertSum model introduced in [36] as our CSS base model. In the fine-tuning CSS model, we use BERT to encode the input text. Specifically, the model splits the entire text into multiple sentences and then passes each sentence as input to BERT. The BERT model, through its deep neural network structure, captures the semantic information in the sentence and transforms it into a vector. This vector is not merely a simple combination of the words in the sentence; it encapsulates the semantic meaning of the sentence within its context. Thus, the model can better understand the importance of each sentence.

$$t_i = BERT(x_i) \tag{3}$$

Here, t_i represents the vector representation of the ith sentence after being processed by BERT, and x_i is the ith sentence.

After obtaining the vector representation t_i for each sentence, the next task is to assign an importance score to each sentence. To achieve this, CSS uses a linear layer. This linear layer functions like a simple "weighted sum" operation, transforming the sentence vector based on the given weight W_h and bias b_h. Then, the model applies a non-linear activation function, tanh, which helps capture more complex relationships.

$$h_i = tanh(W_h t_i + b_h) \tag{4}$$

Here, h_i is the score of the ith sentence, and W_h and b_h are the weight and bias parameters of the linear layer, respectively. By applying the tanh function, the model can better recognize the non-linear features of each sentence, thereby assigning a more accurate score to each sentence.

After obtaining the score for each sentence, the next step is to convert these scores into probabilities. In this step, the CSS model uses the Softmax function to convert a set of scores (e.g., scores of multiple sentences) into a probability distribution. Specifically, Softmax calculates the relative probability of each sentence based on the scores of all sentences. The formula is as follows:

$$p_i = softmax(W_p h_i + b_p) \tag{5}$$

Here, p_i is the probability of the ith sentence, representing the likelihood of this sentence being selected for the summary. W_p and b_p are the weight and bias parameters of another linear layer. Based on the probability p_i of each sentence, the model selects the sentence with the highest probability as a candidate for generating the summary. After selecting these sentences, they ultimately form the main content of the text.

The set of images after the TIFS module and their corresponding textual descriptions are formally defined as follows: $I = \left\{P^i\right\}_{i=1}^m$ and $T = \left\{w^i\right\}_{i=1}^k$, where p^i represents the ith padding of the image, and w^i represents the i th word. The images are divided into m patches, and the text consists of k words. Both the images and text are processed by visual and textual encoders, such as ViT [34] and BERT [37], respectively. The embeddings are defined as $e_P \in R^{m\times C}$ and $e_w \in R^{k\times C}$, where C denotes the number of channels.

To capture informative content, we employ a cross-modal attention module to implicitly model interactions between the image and text modalities. Specifically, we embed $e_P \in R^{m\times C}$ and $e_w \in R^{k\times C}$ as examples. The relationship matrix $R \in R^{m\times k}$ is constructed using matrix multiplication and then passed through a convolutional layer for feature extraction.

$$R = Conv\left(e_P \cdot e_w^T\right) \tag{6}$$

Here, T represents transposition, and $Conv$ is implemented with two convolutional layers. A higher value of R indicates stronger correlation. For the visual modality, we add the text labels to generate an attention vector $v_P \in R^m$. Similarly, we construct a word attention vector $v_w \in R^k$ using the same approach. The vectors v_P and v_w are then aggregated into e_P and e_w via channel-wise multiplication and sigmoid activation. The visual embeddings can be represented as follows:

$$e_{ap} = sigmoi\,d\left(v_p\right) \cdot e_P \tag{7}$$

e_{aw} is processed in the same way. The adjusted embedding representation is passed to SID and SSC.

3.3 Semantic Intensified Distribution Modeling

Counterfactual reasoning is crucial in perceiving sarcasm. Some objects or events in the text and image may exist but are not always related to sarcasm. To identify the sarcasm-related content, we introduce a channel weighting strategy to help the model focus on the information truly related to sarcasm.

We use a channel weighting approach to learn sarcasm-related feature representations. This method gradually activates sarcasm-related features by reweighting the relevant information in both images and text. The formula is as follows:

$$r_p = e_{ap} \cdot \sigma\left(ReLU\left(FC\left(e_{ap}\right)\right)\right) \tag{8}$$

Here, r_p represents the re-weighted image embeddings, and σ denotes the channel-wise variance. The text embeddings r_w are processed in a similar manner. After obtaining these region-specific semantic embeddings, we compute the similarity distribution between positive and negative samples and determine their likelihood of belonging to the same category. Specifically, for r_p and r_w, we aggregate them into $\gamma_v \in R^C$ and $\gamma_t \in R^C$, which are calculated as the mean of all patches and word embeddings, respectively. During training, we maintain two memory banks $M_s = \left\{\left(r_v^i, r_t^i\right)\right\}_{i=1}^q$ and $M_{NS} = \left\{\left(r_v^i, r_t^i\right)\right\}_{i=1}^q$ for positive and negative samples, where q denotes the batch size from previous iterations. By observing the data distribution, we adopt a Gaussian distribution to model the mean and variance of similarity scores. The formula is as follows:

$$\mu = \sum_{i=1}^{q} Sim\left(r_v^i, r_t^i\right) \tag{9}$$

$$\sigma = \sqrt{\sum_{i=1}^{q}\left(Sim\left(r_v^i, r_t^i\right) - \mu\right)^2} \tag{10}$$

Here, Sim denotes the cosine similarity function, and μ and σ represent the mean and variance of the Gaussian distribution, respectively. The distributions D_S and D_{nS} are indicated as $D_S \in N\left(\mu_s, \sigma_s\right)$ and $D_{nS} \in N\left(\mu_{ns}, \sigma_{ns}\right)$. We calculate the likelihood of samples belonging to D_S or D_{nS} based on the probability density function.

Subsequently, the model computes the probability of samples belonging to positive or negative classes based on the probability density function of the Gaussian distribution. The formula is as follows:

$$P = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\tau}\left(\frac{sim\left(r_v^i, r_t^i\right) - \mu}{\sigma}\right)^2 \tag{11}$$

Here, τ controls the importance of σ in temperature scaling.

Finally, we calculate the inconsistency λ_{SID}, defined as $P_S - P_{nS}$, to guide the model. This method is more flexible than using a fixed or adaptive threshold, as it leverages the Gaussian distribution to provide a smooth probability, thereby avoiding biases introduced by hard decision-making.

3.4 Siamese Sentiment Contrastive Learning

In multimodal sarcasm detection, sentiment information is crucial. To better capture sentiment inconsistency, we introduce sentiment knowledge into the model. Specifically, the model utilizes a pre-trained sentiment model based on Twitter data, which provides a sentiment polarity score for each word. If the model is unable to provide sentiment information for a specific word, its sentiment value is set to zero.

First, the word embeddings from the self-attention module are input into a "Siamese Layer." This layer consists of a projection head and a classifier, which extract the sentiment embeddings and predict the polarity of each word. Then, the mean squared error (MSE) is computed for each word's sentiment loss. The formula is as follows:

$$L^{tS} = \frac{1}{k}\sum_{1=1}^{k}\left(\widehat{P}_w^i - P_w^i\right)^2 \tag{12}$$

Here, P_w^i represents the polarity of the ith word.

In the previous cross-attention module, embeddings were aligned, and we used a shared parameter projection head and classifier to process the sentiment representation of the images. Specifically, similar to SID, we obtain the mean of patch and word embeddings as the visual and textual [CLS] representations. Then, $\widehat{P_v} \in R^B$ and $\widehat{P_t} \in R^B$ represent the sentiment polarities of the entire batch, where B is the mini-batch size. To enhance the sentiment representation of the images, we introduce a contrastive learning strategy with continuous supervision labels to capture sentiment polarity strength. A supervised loss G_p is constructed to measure the sentiment polarity discrepancy between images and text. If the discrepancy is large, the embeddings are pushed apart; if the discrepancy is small, the embeddings are pulled closer. The formula is as follows:

$$G_p^{ij} = softma\left(\exp\left(-\left|\widehat{P}_v^i - \widehat{P}_t^i\right|\right)\right) \tag{13}$$

Here, P_v^i and P_t^i denote the polarity strengths, ranging from [−1, 1]. The similarity matrix G_e can be obtained by calculating the dot product between the sentiment embeddings $S_v \in R^{B\times C}$ and $S_t \in R^{B\times C}$, which are outputs from the projection head.

$$G_e^{ij} = softma\left(\exp\left(s_v^i, s_t^i\right)\right) \tag{14}$$

Additionally, the loss for continuous After the SID and SSC modules, the embeddings undergo experimental fusion for the final prediction. For embeddings from the same modality, element-wise multiplication is performed, followed by concatenation from both the semantic and sentiment perspectives.

Contrastive learning is calculated using the Kullback-Leibler (KL) divergence as follows:

$$L^{cc} = kL\left(G_e, G_p\right) \tag{15}$$

We define the sentiment polarity discrepancy between the visual and textual modalities as another factor for sarcasm detection, represented as $\lambda ssc = |P_v - P_t|$.

Considering the inconsistencies in factual and sentiment layers, we integrate the predicted embeddings y_f with two distinct inconsistency factors, λ_{SID} and λ_{SSC}.

$$\widehat{y} = sigmoid\left(y_f + \lambda_{SID} + \lambda_{SSC}\right) \tag{16}$$

The binary cross-entropy loss is calculated as:

$$L^{bce} = -\left[y \cdot log\left(\widehat{y}\right) + (1 - y) \cdot \log\left(1 - \widehat{y}\right)\right] \tag{17}$$

Finally, the PKME-MLM network for multimodal sarcasm detection is optimized through the computed loss:

$$L = L^{bce} + L^{cc} + L^{tS} \tag{18}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

We conducted experiments on a public MSD dataset [38], which consists of 24,635 tweets divided into a training set (19,816 tweets), a validation set (2410 tweets), and a test set (2409 tweets). Each tweet is labeled as sarcastic (positive) or non-sarcastic (negative), and the dataset contains both text and image-text pairs.

The text in the dataset is primarily in English and contains various expressions of sarcasm and non-sarcasm. For preprocessing, we cleaned the data by removing offensive content, URLs, and irrelevant words. The dataset is split into a training set, validation set, and test set, accounting for 80%, 10%, and 10% of the total data, respectively. We used accuracy, precision, recall, binary average, and macro average as evaluation metrics.

4.2 Quantitative Analysis

Through extensive experimental research, we conducted a comprehensive evaluation of the application of multi-level perception networks in sarcasm detection tasks and compared their performance with various existing unimodal (image and text) and multimodal models in sarcasm detection. The results are presented in Table 1 (This table lists the performance of various unimodal (image and text) and multimodal models in sarcasm detection tasks, with evaluation metrics including accuracy, binary average precision, recall, and macro average F1 score.), leading to the following analyses and observations:

(1) The comparative experiments reveal that our model, PKME-MLM, achieves the highest accuracy, reaching 94.35%. This indicates that, with the incorporation of the TIFS and Twitter-specific models, our model attains leading performance in sarcasm recognition tasks, significantly surpassing other multimodal and unimodal models.

(2) In the binary classification and macro average metrics, the PKME-MLM model achieves a binary average precision of 93.07% and a macro average recall of 91.70%. This indicates that the model not only demonstrates high precision in classification but also maintains relatively balanced performance across

different sentiment categories, showcasing higher stability and reliability. Compared to other models, such as DIP and HKE, PKME-MLM excels across all metrics, highlighting its stronger capabilities.

(3) These experimental results support the important notion that in multimodal tasks where both images and texts are involved, the performance of the PKME-MLM model surpasses that of other multimodal models, such as DIP and CMGCN. This indicates that the introduction of TIFS technology and the Twitter-specific model significantly enhances the integration capability of multimodal information, enabling the model to better capture the complex relationships between text and images, thereby improving overall sarcasm detection performance.

Table 1: Comparison of accuracy and evaluation metrics for different models in sarcasm recognition tasks

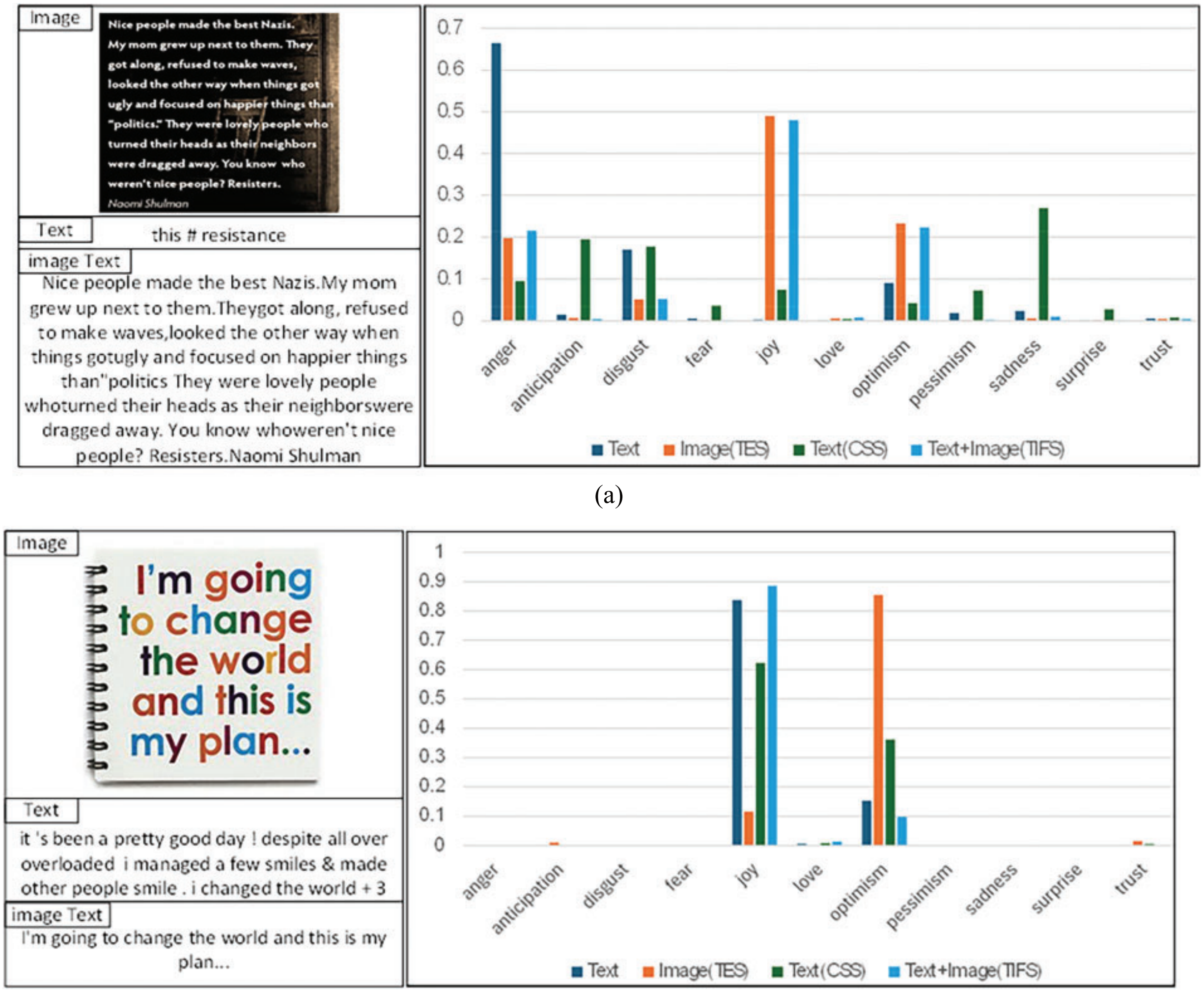
Modality	**Method**	**Acc.**	**Binary-average**			**Macro-average**		
			Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Image	ResNet [38]	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT [34]	67.83	57.93	70.07	63.43	65.68	71.35	68.40
Text	Bi-LSTM [39]	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN [40]	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD [41]	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT-Base [37]	83.85	78.72	82.27	80.22	81.31	80.97	81.90
Image+Text	HFM [38]	86.63	83.84	84.18	84.01	86.24	86.28	86.26
	InCrossMGs [42]	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	HKE [43]	87.36	81.84	86.48	84.09	–	–	–
	CMGCN [44]	87.55	81.63	84.69	84.16	87.02	86.97	87.00
	DIP [45]	89.59	87.76	86.56	87.17	88.46	89.13	89.01
	PKME-MLM (Our)	94.35	93.07	91.70	92.38	93.92	94.21	94.31

4.3 Qualitative Analysis

In our study, to evaluate the performance of the PKME-MLM model in multimodal sarcasm detection, we analyzed the effects of different input types under various modules (such as text-only input, text extracted from images, text summary from CSS, and image-text fusion from TIFS) on capturing sentiment polarity values across different sentiment categories, as shown in Fig. 4. These sentiment categories include anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. The figure compares the performance of different input types (such as text-only input, text extracted from images, text summary from CSS, and image-text fusion from TIFS) across different sentiment dimensions. The detection performance of different modules on sentiment polarity values is illustrated using bar charts.

These qualitative experiments allowed us to gain deeper insights into the roles of the TIFS module and the Twitter-specific module in the PKME-MLM model and to verify their effectiveness in enhancing multimodal sarcasm detection capabilities.

In Fig. 4a, the PKME-MLM module demonstrates a significant polarity value of 0.6649 in the "anger" dimension with text-only input (Text). This indicates that PKME-MLM can accurately identify and quantify strong negative emotions in the text, particularly the anger often found in sarcastic content. In contrast, the image-text fusion strategy (TIFS) shows a more balanced performance in the "anger" dimension, with a polarity value of 0.2152. By incorporating image information, the TIFS module reduces the extreme emotions conveyed by text-only input, demonstrating a more comprehensive sentiment analysis capability. This balance reflects the advantage of TIFS in capturing complex sarcastic emotions, allowing the model to rely not only on strong negative emotions in the text but also to understand emotional diversity more precisely through multimodal information.



Nice people made the best Nazis. My mom grew up next to them. They got along, refused to make waves, looked the other way when things got ugly and focused on happier things than "politics." They were lovely people who turned their heads as their neighbors were dragged away. You know who weren't nice people? Resisters.
Naomi Shulman

this # resistance

Nice people made the best Nazis.My mom grew up next to them.Theygot along, refused to make waves,looked the other way when things gotugly and focused on happier things than"politics They were lovely people whoturned their heads as their neighborswere dragged away. You know whoweren't nice people? Resisters.Naomi Shulman

(a)

it 's been a pretty good day ! despite all over overloaded i managed a few smiles & made other people smile . i changed the world + 3

I'm going to change the world and this is my plan...

(b)

Figure 4: The recognition results of the PKME-MLM model for eleven sentiment categories in the multimodal sarcasm detection task

In Fig. 4b, the PKME-MLM module performs well in capturing positive emotions with text and image inputs. The polarity value for text input in the "joy" dimension reaches 0.8367, while the polarity value for image input in the "optimism" dimension is 0.8548, showcasing the PKME-MLM module's sensitivity in detecting positive emotions. However, the TIFS module, by combining text and image, demonstrates a more balanced performance across emotional dimensions, particularly in the subtler "love" emotion, with a polarity value of 0.0079. Although this polarity value is relatively low, it shows that TIFS effectively captures more complex emotional expressions in both text and images through multimodal fusion, highlighting its stability and broad applicability in handling multi-emotion tasks.

These qualitative experiments demonstrate the effectiveness of the PKME-MLM module in multi-emotion polarity analysis across eleven emotional dimensions, accurately distinguishing and quantifying complex emotions in sarcastic text. Moreover, the TIFS module, by leveraging the advantages of fusing text and image information, shows outstanding performance in multimodal sarcasm detection, especially in dealing with emotional complexity and subtle emotional expressions. TIFS not only enhances the model's balance between positive and negative emotions but also provides a more comprehensive emotion recognition capability when handling sarcastic texts.

4.4 Ablation Study

To investigate the effectiveness of the components of the TIFS and Twitter modules in the multi-modal sarcasm recognition task, we designed a series of ablation experiments. In these experiments, we first assessed the impact of integrating the TIFS module alone to validate its capability in fusing image and text information. Subsequently, we evaluated the effect of introducing the improved Twitter-specific module independently to explore its advantages in processing social media texts. Finally, we assessed the combined effect of both modules to verify their synergistic impact in multi-label sentiment recognition tasks, demonstrating a significant enhancement in our model's ability to handle complex emotional expressions. The experimental results are presented in Table 2.

Table 2: Ablation experiments

Method	Acc.	Binary-average			Macro-average		
		Precision	Recall	F1	Precision	Recall	F1
TIFS	94.02	91.73	91.44	91.58	93.60	93.64	93.62
Twitter	93.95	93.66	88.83	91.18	93.90	93.12	93.46
PKME-MLM	94.35	93.07	91.70	92.38	93.92	94.21	94.31

Based on the results presented in Table 2, we have the following observations:

(1) The TIFS module performs exceptionally well in both binary average and macro average, especially with a macro F1 score of 91.58, demonstrating its strong capability in fusing image and text information. This indicates that the TIFS module can stably capture different categories of sentiment, particularly showing good balance in multi-modal sentiment recognition tasks.

(2) The Twitter module excels in binary average precision, reaching 93.66, showing its high accuracy in processing social media texts. However, the macro recall rate is slightly lower, especially with a macro F1 of 91.18, indicating some shortcomings in covering sentiment categories, which may be related to the complexity and diversity of social media texts.

(3) The PKME-MLM model performs excellently across all metrics, particularly leading in macro recall (94.21) and macro F1 (94.31), showcasing the synergistic effect after combining the TIFS and Twitter modules. Compared to the individual TIFS or Twitter modules, PKME-MLM demonstrates more comprehensive and stable performance, indicating that the combination of multi-modal fusion and social media-specific processing brings significant improvements in sentiment recognition tasks.

5 Conclusions

Due to the issues of insufficient information utilization and simplified sentiment labels in multimodal sarcasm detection, models face limitations in accurately capturing complex sarcastic expressions. To address these challenges, we propose an innovative approach. First, we designed the Text-Image Fusion Summary (TIFS) model, which extracts textual information from images and fuses it with the original text, generating a more comprehensive text sequence. This approach not only expands the sources of textual data but also enhances semantic richness, improving the accuracy of sarcasm detection.

Additionally, we introduced a Twitter-specific model, fine-tuned on a large-scale Twitter dataset, using a multi-label classification approach that allows tweets to have multiple sentiment labels simultaneously. This design enriches sentiment recognition and improves sarcasm detection accuracy.

These models significantly enhance sarcasm recognition performance, marking an important advancement in the field. However, the potential limitations of this study include its reliance on large-scale labeled datasets and the model's generalization issues in other domains or languages. In particular, there are challenges in the model's ability to generalize when handling diverse and complex sarcastic expressions. In the future, we plan to introduce more diverse datasets and explore cross-lingual sarcasm detection to enhance the model's adaptability in different languages and cultural contexts. Additionally, we will integrate more dynamic and context-aware methods to improve the model's robustness in real-time applications.

Acknowledgement: This work was supported in part by the Laboratory for Artificial Intelligence and International Communication.

Funding Statement: This work was funding partly by the National Natural Science Foundation of China under grant number 61701179.

Author Contributions: All authors participated in designing and implementing the study. All authors discussed the basic structure of the manuscript. Jian Luo and Yaling Li drafted the main parts of the manuscript. Xuliang Hu reviewed and edited the draft. Yaling Li, Xueyu Li and Xuliang Hu participated in the experiments. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in (MSD dataset) at https://github.com/ZLJ2015106/pytorch-multimodal_sarcasm_detection.git (accessed on 01 January 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Liu H, Wei R, Tu G, Lin J, Liu C, Jiang D. Sarcasm driven by sentiment: a sentiment-aware hierarchical fusion network for multimodal sarcasm detection. Inf Fusion. 2024;108(1):1–10. doi:10.1016/j.inffus.2024.102353.
2. Gibbs RW. On the psycholinguistics of sarcasm. J Exp Psychol: Gen. 1986;115(1):3–15. doi:10.1037/0096-3445.115.1.3.
3. Rajadesingan A, Zafarani R, Liu H. Sarcasm detection on twitter: a behavioral modeling approach. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; 2015 Jan 31–Feb 06; Shanghai, China. New York, NY, USA: Association for Computing Machinery; 2015. p. 97–106.
4. Babanejad N, Davoudi H, An A, Papagelis M. Affective and contextual embedding for sarcasm detection. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020 Dec 08–13; Barcelona, Spain Stroudsburg, PA: International Committee on Computational Linguistics; 2020. p. 225–43.
5. Ashwitha A, Shruthi G, Shruthi HR, Upadhyaya M, Ray AP, Manjunath TC. Sarcasm detection in natural language processing. Mater Today: Proc. 2021;37(2):3324–31. doi:10.1016/j.matpr.2020.09.124.

6. Zhang Y, Liu Y, Li Q, Tiwari P, Wang B, Li Y, et al. CFN: a complex-valued fuzzy network for sarcasm detection in conversations. IEEE Trans Fuzzy Syst. 2021;29(12):3696–710. doi:10.1109/TFUZZ.2021.3072492.
7. Amir S, Wallace BC, Lyu H, Carvalho P, Silva MJ. Modelling context with user embeddings for sarcasm detection in social media. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning; Berlin, Germany; 2016. p. 167–77. doi:10.48550/arXiv.1607.00976.
8. Wu Y, Zhao Y, Lu X, Qin B, Wu Y, Sheng J, et al. Modeling incongruity between modalities for multimodal sarcasm detection. IEEE Multimed. 2021;28(2):86–95. doi:10.1109/MMUL.2021.3069097.
9. Ghosh A, Veale T. Magnets for sarcasm: making sarcasm detection timely, contextual and very persona. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep 07–11; Copenhagen, Denmark. Stroudsburg, PA: Association for Computational Linguistics; 2017. p. 482–91.
10. Kannangara S. Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining; 2018 Feb 05–09; Marina Del Rey, CA, USA. New York, NY, USA: Association for Computing Machinery; 2018. p. 751–2.
11. Liu P, Chen W, Ou G, Wang T, Yang D, Lei K. Sarcasm detection in social media based on imbalanced classification. In: Web-Age Information Management: 15th International Conference; 2014 Jun 16–28; Macau, China. Heidelberg, Germany: Springer International Publishing; 2014. p. 459–71.
12. Barbieri F, Camacho-Collados J, Neves L, Espinosa-Anke L. TweetEval: unified benchmark and comparative evaluation for tweet classification. In: Findings of the association for computational linguistics: EMNLP 2020; Stroudsburg, PA: Association for Computational Linguistics; 2020. p. 1644–50. doi:10.48550/arXiv.2010.12421.
13. Riloff E, Qadir A, Surve P, De Silva L, Gilbert N, Huang R. Sarcasm as contrast between a positive sentiment and negative situation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; 2013 Oct 18–21; Seattle, WA, USA. Stroudsburg, PA: Association for Computational Linguistics; 2013. p. 704–14.
14. Joshi A, Bhattacharyya P, Carman MJ. Automatic sarcasm detection: a survey. ACM Comput Surv. 2017;50(5):1–22. doi:10.1145/3124420.
15. Mukherjee S, Bala PK. Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering. Technol Soc. 2017;48:19–27. doi:10.1016/j.techsoc.2016.10.003.
16. Sarsam SM, Al-Samarraie H, Alzahrani AI, Wright B. Sarcasm detection using machine learning algorithms in Twitter: a systematic review. Int J Mark Res. 2020;62(5):578–98. doi:10.1177/1470785320921779.
17. Porwal S, Ostwal G, Phadtare A, Pandey M, Marathe MV. Sarcasm detection using recurrent neural network. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS); 2018 Jun 14–15; Madurai, India. Piscataway, NJ, USA: IEEE; 2018. p. 746–8.
18. Salim SS, Ghanshyam N, Agrawal A, Darkunde M, Mazahir B, Dungarpur B, et al. Deep LSTM-RNN with word embedding for sarcasm detection on twitter. In: 2020 International Conference for Emerging Technology (INCET); 2020 Jun 05–07; Belgaum, India. Piscataway, NJ, USA: IEEE; 2020. p. 1–4.
19. Kumar A, Narapareddy VT, Aditya Srikanth V, Malapati A, Neti LBM. Sarcasm detection using multi-head attention based bidirectional LSTM. IEEE Access. 2020;8(8):6388–97. doi:10.1109/ACCESS.2019.2963630.
20. Joshi A, Sharma V, Bhattacharyya P. 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 Jul 26–31; Beijing, China. Stroudsburg, PA, USA: Association for Computational Linguistics; 2015. p. 757–62.
21. Ding N, Tian S, Yu L. A multimodal fusion method for sarcasm detection based on late fusion. Multimed Tools Appl. 2022;81(6):8597–616. doi:10.1007/s11042-022-12122-9.
22. Yue T, Mao R, Wang H, Hu Z, Cambria E. KnowleNet: knowledge fusion network for multimodal sarcasm detection. Inf Fusion. 2023;100:101921. doi:10.1016/j.inffus.2023.101921.
23. Achlioptas P, Ovsjanikov M, Haydarov K, Elhoseiny M, Guibas LJ. Artemis: affective language for visual art. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Nashville, TN, USA. Piscataway, NJ, USA: IEEE; 2021. p. 11569–79.

24. Yaakub MR, Latiffi MIA, Zaabar LS. A review on sentiment analysis techniques and applications. In: IOP Conference Series: Materials Science and Engineering; 2019 Feb 04–05; Bangkok, Thailand. Bristol, UK: IOP Publishing Ltd.; 2019. Vol. 551, No. 1, p. 1–4.
25. Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10); 2010 May 17–23; Valletta, Malta. Paris, France: European Language Resources Association; 2010. p. 2200–4.
26. Zainuddin N, Selamat A. Sentiment analysis using support vector machine. In: 2014 International Conference on Computer, Communications, and Control Technology (I4CT); 2014 Sep 02–04; Langkawi, Malaysia. Piscataway, NJ, USA: IEEE; 2014. p. 333–7.
27. Parmar H, Bhanderi S, Shah G. Sentiment mining of movie reviews using random forest with tuned hyperparameters. In: International Conference on Information Science; 2014 Jul 04; Kerala, India. Piscataway, NJ, USA: IEEE; 2014. p. 1–6.
28. Goel A, Gautam J, Kumar S. Real time sentiment analysis of tweets using Naive Bayes. In: 2016 2nd International Conference on Next Generation Computing Technologies (NGCT); 2016 Oct 14–16; Dehradun, India. Piscataway, NJ, USA: IEEE; 2016. p. 257–61.
29. Liao S, Wang J, Yu R, Sato K, Cheng Z. CNN for situations understanding based on sentiment analysis of twitter data. Procedia Comput Sci. 2017;111:376–81. doi:10.1016/j.procs.2017.06.037.
30. Li D, Qian J. Text sentiment analysis based on long short-term memory. In: First IEEE International Conference on Computer Communication and the Internet (ICCCI); 2016 Oct 13–15; Wuhan, China. Piscataway, NJ, USA: IEEE; 2016. p. 471–75.
31. Naseem U, Razzak I, Musial K, Imran M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. Future Gener Comput Syst. 2020;113:58–69. doi:10.1016/j.future.2020.06.050.
32. Liu N, Zhao J. A BERT-based aspect-level sentiment analysis algorithm for cross-domain text. Comput Intell Neurosci. 2022;2022:8726621. doi:10.1155/2022/8726621.
33. Liao W, Zeng B, Yin X, Wei P. An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa. Appl Intell. 2021;51(6):3522–33. doi:10.1007/s10489-020-01964-1.
34. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. 2020. doi:10.48550/arXiv.2010.11929.
35. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning; 2021 Jul 18–24; San Francisco, CA, USA: PMLR; 2021. p. 10347–57.
36. Liu Y. Fine-tune BERT for extractive summarization. 2019. doi:10.48550/arXiv.1903.10318.
37. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. doi:10.48550/arXiv.1810.04805.
38. Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in twitter with hierarchical fusion mode. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 02; Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 2506–15.
39. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
40. Tay Y, Tuan LA, Hui SC, Su J. Reasoning with sarcasm by reading in-between. 2018. doi:10.48550/arXiv.1805.02856.
41. Xiong T, Zhang P, Zhu H, Yang Y. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In: The World Wide Web Conference; 2019 May 13–17; San Francisco, CA, USA. New York, NY, USA: Association for Computing Machinery; 2019. p. 2115–24.
42. Liang B, Lou C, Li X, Gui L, Yang M, Xu R. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; China. New York, NY, USA: Association for Computing Machinery; 2021. p. 4707–15.
43. Liu H, Wang W, Li H. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. 2022. doi:10.48550/arXiv.2210.03501.

44. Liang B, Lou C, Li X, Yang M, Gui L, He Y, et al. 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 May 22–27; Dublin, Ireland. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022. p. 1767–77.

45. Wen C, Jia G, Yang J. Dip: dual incongruity perceiving network for sarcasm detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 18–22; Vancouver, BC, Canada. Piscataway, NJ, USA: IEEE; 2023. p. 2540–50.