Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050384

ARTICLE

Research on Sarcasm Detection Technology Based on Image-Text Fusion

Xiaofang Jin[1], Yuying Yang[1,*], Yinan Wu[1] and Ying Xu[2]

[1]School of Information and Communication Engineering, Communication University of China, Beijing, 100024, China

[2]Cable Television Technology Research Institute, Academy of Broadcasting Science, Beijing, 100053, China

*Corresponding Author: Yuying Yang. Email: yyy_cuc@163.com

Received: 05 February 2024 Accepted: 20 May 2024 Published: 20 June 2024

ABSTRACT

The emergence of new media in various fields has continuously strengthened the social aspect of social media. Netizens tend to express emotions in social interactions, and many people even use satire, metaphors, and other techniques to express some negative emotions, it is necessary to detect sarcasm in social comment data. For sarcasm, the more reference data modalities used, the better the experimental effect. This paper conducts research on sarcasm detection technology based on image-text fusion data. To effectively utilize the features of each modality, a feature reconstruction output algorithm is proposed. This algorithm is based on the attention mechanism, learns the low-rank features of another modality through cross-modality, the eigenvectors are reconstructed for the corresponding modality through weighted averaging. When only the image modality in the dataset is used, the preprocessed data has outstanding performance in reconstructing the output model, with an accuracy rate of 87.6%. When using only the text modality data in the dataset, the reconstructed output model is optimal, with an accuracy rate of 85.2%. To improve feature fusion between modalities for effective classification, a weight adaptive learning algorithm is used. This algorithm uses a neural network combined with an attention mechanism to calculate the attention weight of each modality to achieve weight adaptive learning purposes, with an accuracy rate of 87.9%. Extensive experiments on a benchmark dataset demonstrate the superiority of our proposed model.

KEYWORDS

Sentiment analysis; sarcasm detection; feature fusion; feature reconstruction

1 Introduction

Benefiting from the development of global science and technology, the Internet business has been rapidly popularized, and various social software have poured into everyone's life. The latest social media trend report published by GlobalWebIndex (GWI) [1] pointed out that starting from 2018, Chinese Internet users spend more than 2 h per day on social media on average, ranking first in the world. Correspondingly, the emotional instability of Chinese Internet users due to social media is increasing year by year, ranking high in the Asia-Pacific region. The "social" attribute in social media has been continuously strengthened, and more and more people tend to vent their negative emotions in social interactions to reduce their mental stress. To address issues arising from excessive freedom of speech, the platform manually filters negative or distorted content. However, this intervention

method is ineffective in managing daily data flow of over 100 million, especially when netizens tend to use sarcasm or metaphors to express negative emotions. As data holds value across all sectors, the government also employs social media comment data for public opinion analysis to formulate reasonable policies and guide public opinion.

Given the nascent state of domestic research on sarcasm detection and the dearth of readily accessible public datasets, this article will continue to rely on the foreign Twitter image-text dual-modal dataset. The Twitter dataset is comparable to the domestic social platform data in all aspects except for the language form, which is English. The natural language processing methods used for Chinese and English are identical, so the Twitter dataset can well represent the information carried by domestic social media. This paper argues that any information posted by users in social media is highly valuable for sarcasm detection research tasks, including topics, symbols, texts, pictures, and deeper modal conflicts.

As illustrated in Fig. 1, the user expresses enthusiasm for spending several hours each day with the books in the coming weeks. The image depicts a large number of books and a crying face emoticon in the lower right corner, which may be interpreted as an allusion to the challenges faced by law students. In this instance, relying solely on a single text message as the data source for detecting sarcasm would be a clear mistake. However, by combining the image message with the accompanying text, it becomes evident that there is a discrepancy between the two, which allows us to conclude that this tweet is sarcastic in nature.

Figure 1: Twitter example image

The main contributions of this paper are divided into two parts:

1. Once the feature vectors for both the image and text data in social media comments have been obtained, the secondary processing can be performed to optimize the experimental effect. This

paper proposes a feature reconstruction output algorithm. The attention mechanism enables the algorithm to learn low-rank features of another modality across modalities, calculate the weight of each original vector, and obtain the reconstructed feature vector of the corresponding modality through weighted average. A comparison with the feature direct output method demonstrates the effectiveness of the feature reconstruction output algorithm in handling text modal data.

2. Effective fusion of image data and text data in social media comment information can improve the performance of sarcasm detection experiments. This paper presents a weight adaptive learning algorithm, which uses neural network combined with attention mechanism to calculate the attention of each mode and achieve the purpose of weight adaptive learning. This method has a positive effect on improving the experimental results of sarcasm detection.

2 Related Work

2.1 Research on Sarcasm Detection with Rule-Based Approach

Gonzalez et al. [2] defined sarcasm as the opposite of the literal and intended meaning in a blog. According to Gonzalez, sarcasm on social media conveys negative sentiment, while literal statements are positive.

In large-scale classification tasks, preprocessing algorithms are a major focus of scientific research. Sparse representation is an optimal approach for image preprocessing as it utilizes less data to represent the complete signal. Zhu et al. [3] proposed an image fusion scheme based on image cartoon texture decomposition and sparse representation. Min et al. [4] employed the L2 norm to maintain the sparsity of the coefficient matrix rows during the minimization of reconstruction error. Sun et al. [5] proposed a structured robust adaptive dictionary learning framework for discriminative sparse representation learning. In general, researchers select a large-scale dictionary in order to ensure optimal representation capability. Nevertheless, this can result in an increased time requirement for the sparse representation process, which in turn may lead to a reduction in experimental efficiency.

2.2 Research on Sarcasm Detection Based on Machine Learning

Following text representation, text data is converted into a vector form that can be directly processed by the computer. When analyzing image data, the process differs from text processing. It involves data preprocessing, feature extraction, and training classification.

The advent of social media platforms has led to the emergence of a popular trend in which satirical effects are achieved through the combination of text and images on Twitter. To address the challenges associated with the detection of sarcasm on social media, Zhao et al. [6] established a multimodal framework that captures both textual and visual information. They proposed a Canyond-Attention Network (CAN) that effectively integrates text and image information into a unified framework. For the details in the data, such as the pitch changes of audio signals, facial expressions and body postures in images, Ding et al. [7] proposed a multimodal learning framework based on residual connections that can effectively perform late fusion. Liang et al. [8] determined the emotional inconsistency within a certain modality and between different modalities using Heterogeneous and Cross-Modal Graphs (InCrossMGs) for each multimodality. They employed the Interactive Graph Convolutional Networks (IGCN) structure is used to jointly and interactively learn the incongruous relations of modality graphs, with the objective of identifying important cues in sarcasm detection.

Internet memes (MemEs) have become a powerful tool for disseminating political, psychological, and sociocultural ideas. Pramanick et al. [9] proposed a network called Multimodal Framework for Detecting Harmful MemEs and Their Targets (MOMENTA) for the purpose of detecting harmful memes and their targets, which use global and local perspectives to detect harmful memes. Shah et al. [10] proposed a method for capturing inconsistencies through subword-level embeddings learned by fastText to address the joint problem of code-mixing and sarcasm detection. Ren et al. [11] proposed the Hierarchical Attention-based Model with Multi-task Self-Training (HAMLET) model for comprehensive analysis of users. The model is based on hierarchical attention and a multi-task self-training algorithm with sparse sharing. Qiu et al. [12] proposed a neural network model based on multi-layer attention. The model employs an improved Latent Dirichlet Allocation (LDA) and paragraph vector learning framework to obtain text vector representations. It captures text context information through the Bidirectional Long Short-Term Memory (Bi-LSTM) layer, and finally classifies and predicts sentiment factors in a Convolutional Neural Network (CNN).

In a multimodal environment from social media, Lucas et al. [13] proposed a classification model based on the transferable learning capabilities of the Contrastive Language-Image Pretraining (CLIP) neural network architecture. Gupta et al. [14] employed a RoBERTa model with joint attention to incorporate input text and image attributes, address contextual inconsistencies between the two, and integrate feature affine transformations by adjusting input image and text features through FiLMed ResNet to capture multimodal information. Kumar et al. [15] proposed Sarcasm Detection in Dialogue (SED), which focuses on the discourse structure of sarcastic dialogue. The objective of this task is to generate natural language explanations for sarcastic dialogues. To this end, a Multimodal Context-Aware Attention and Global Information Fusion Module (MAF) has been proposed, which is designed to interpret the latent sarcastic connotations present in dialogues.

Sarcasm detection research frequently examines syntactic, lexical, or pragmatic features expressed through words, emoji, and exclamation points. Kumar et al. [16] utilized FastText word embeddings in combination with a Bi-directional Encoder Representations from Transformers (BERT) language model to identify these text features. Savini et al. [17] proposed a migration learning model for sarcasm detection based on BERT pre-training, which can use emotion classification and emotion detection as a separate intermediate task tot inject feature knowledge into the target task of sarcasm detection. Tashu et al. [18] proposed a multimodal emotion recognition architecture that uses both feature-level attention and modality attention to classify emotions in art. Pramanick et al. [19] proposed a multimodal learning system for optimal transmission, which uses self-attention to learn the optimal transmission within and across modalities. Additionally, it also fuses the attention between each modality and combines multimodal mechanisms to capture the interdependencies between the modalities.

2.3 Attention Mechanism

Similar to the visual attention mechanism of the brain, the eyeball first scans the global information in the visual field, and the central nervous system of the brain reacts to obtain the image target area that focuses on, and then the eyeball scans the area in detail to obtain the local area within the target range information, while selectively ignoring the information outside the area. The attention mechanism in the neural network is to allocate computing resources obliquely and prioritizes important tasks when the computing power of the system is limited. This allows for efficient identification of information beneficial to the current analysis task from massive data. Usually, the more network model parameters, the easier it is for the model to learn features. However, this also results in a larger amount of information being stored in the network model, which may lead to

the problem of information overload. The introduction of the attention mechanism allows model to focus on researching the input information for the current task, reducing its sensitivity to information outside the region and solving the problem of information overload.

As an algorithmic idea, the attention mechanism plays an important role in many fields [20–24]. It is commonly employed in combination with the Encoder-Decoder framework, as shown in Fig. 2.

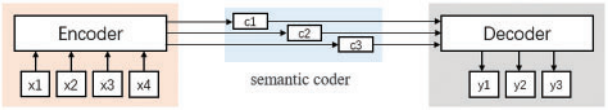

Figure 2: Encoder-Decoder framework

3 Our Approach

This section is divided into two parts. The initial section presents the secondary processing methodology for the two distinct types of feature vectors, following the model's acquisition of the feature vectors associated with the input image and text modalities. The method of directly outputting the image feature vector and text feature vector to the modal fusion part without processing is called feature direct output. An alternative method to this is feature reconstruction output, which is based on the attention mechanism and cross-modality to learn the low-rank features of another modality, calculates the weight of each original vector, and obtain the reconstructed feature vector of the corresponding modality through weighted average. The second part describes the modality fusion method based on the sarcasm detection task. The weight ratio fusion method treats each modality separately, outputs them individually, and finally adds a certain weight to each output to form the final result. The second method involves building a neural network to learn the weights of the component vectors of the image-text modality, resulting in a one-dimensional fusion vector. This method is referred to as weight adaptive learning.

Fig. 3 shows the positions of the feature reconstruction output algorithm and feature fusion algorithm in the sarcasm detection research framework.

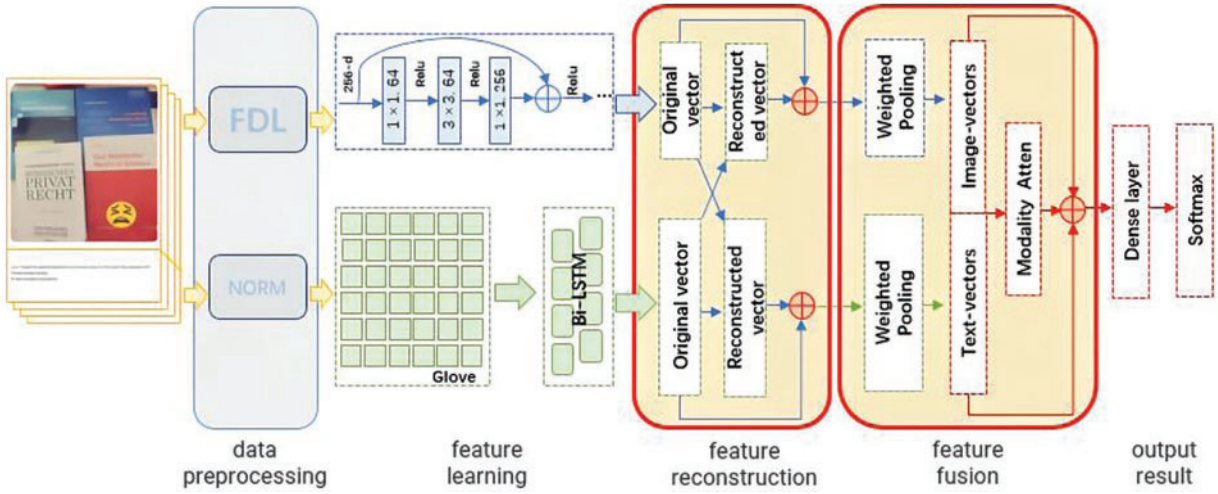

Figure 3: Schematic diagram of the location of the feature reconstruction output algorithm

3.1 Feature Reconstruction Output Algorithm

In the context of sarcasm detection tasks based on image-text fusion, it is important to recognise that a single modality may not always convey all the intended information. Consequently, it is crucial to leverage the distinctions between the image and text modalities to enhance the characteristics of each modality. The concept of attention mechanisms inspired the reconstruction of features after the original image and text feature vectors were learned. This process generated image and text vectors containing cross-modality features. The schematic diagram of the feature reconstruction output is presented in Fig. 4. The difference between this and the feature direct output method is reflected in the red frame area.

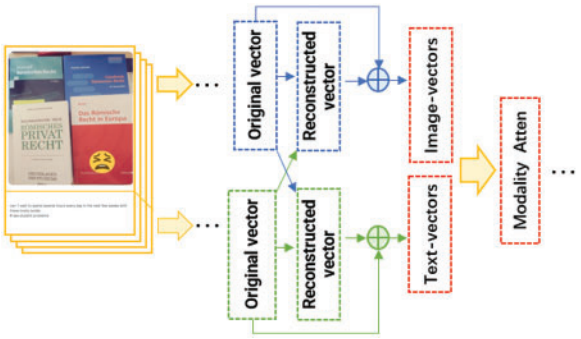

Figure 4: Schematic diagram of feature reconstruction output

The weight corresponding to the i-th original vector of modality p under the guidance of modality q is $\boldsymbol{\alpha}_{pq}^{(i)}$, which is given by:

$$\boldsymbol{\alpha}_{pq}^{(i)} = \boldsymbol{W}_{pq1} \tanh\left(\boldsymbol{W}_{pq2}\left[\boldsymbol{X}_{p}^{(i)}, \boldsymbol{V}_{q}\right] + \boldsymbol{b}_{pq2}\right) + \boldsymbol{b}_{pq1} \tag{1}$$

where $p, q \in \{image, text\}_{modality}$, i is the number of original vectors contained in each original modality, $\boldsymbol{X}_{p}^{(i)}$ represents the i-th original vector of modality p, $\boldsymbol{W}_{pq1}$ is the weight matrix of modality p under the guidance of modality $q1$, $\boldsymbol{W}_{pq2}$ is the weight matrix of modality p under the guidance of modality $q2$; $\boldsymbol{b}_{pq1}$ and $\boldsymbol{b}_{pq2}$ are the corresponding expected deviations. Aggregating all the original vectors yields:

$$\boldsymbol{\alpha}_{pq} = \text{softmax}\left(\boldsymbol{\alpha}_{pq}\right) \tag{2}$$

The corresponding weight of the i-th original vector of modality p is:

$$\boldsymbol{\alpha}_{p}^{(i)} = \frac{\sum_{q} \boldsymbol{\alpha}_{pq}^{(i)}}{n} \tag{3}$$

n represents the number of modalities, where $n = 2$. In summary, the final feature reconstruction vector is:

$$\boldsymbol{V}_{p} = \sum_{i=1}^{\boldsymbol{L}_{p}} \boldsymbol{\alpha}_{p}^{(i)} \boldsymbol{X}_{p}^{(i)} \tag{4}$$

where $\boldsymbol{L}_{p} = \text{length}\left\{\boldsymbol{X}_{p}^{(i)}\right\}$.

In contrast to the feature reconstruction output algorithm, the feature direct output algorithm obtains the corresponding image feature vector and text feature vector after the model has learned the input data. The two modality vectors are not subjected to further processing and are directly output to the modality fusion component, which ultimately outputs a single fusion vector.

3.2 Feature Fusion Algorithm

The objective of the weight ratio fusion algorithm is to derive the image vector and text vector of the input information. Subsequently, the aforementioned vectors are passed through a fully connected layer of the same size, which outputs the classification result through the softmax function. This classification result is then assigned a weight artificially, and the final classification is obtained through the weight ratio fusion result. The schematic diagram is depicted in Fig. 5.

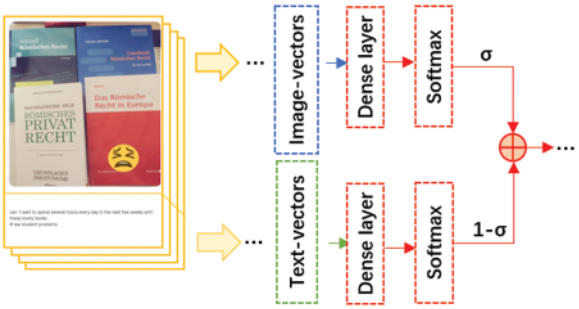

Figure 5: Schematic diagram of weight ratio fusion

In the absence of exceptional circumstances, artificially specifying a certain modality weight is unlikely to yield the optimal effect of modality fusion. Consequently, it is of paramount importance to identify a methodology that can facilitate the organic fusion of multiple modalities. In order to represent the semantic meaning of sarcasm, textual features are employed, while image features are utilized to capture visual cues related to sarcasm. In this manner, information derived from multiple modalities can be integrated to enhance the capacity for sarcasm detection. The use of multiple modalities can provide diverse perspectives and entry points for the understanding of sarcasm, thereby enhancing the accuracy and comprehensiveness of emotion recognition in comparison to monomodality.

As illustrated in Fig. 6, using the neural network combined with the attention mechanism, the attention weight of each modality can be calculated to achieve the purpose of weight adaptive learning. This algorithm is therefore called the weight adaptive learning algorithm.

To align features between modalities, it is necessary to convert the eigenvectors of each modality to the same length. As shown in the following formula, $\boldsymbol{V}_m$ represents the eigenvector of modality m, $\boldsymbol{W}_{m1}$ is a fixed-length weight matrix, $\boldsymbol{b}_{m1}$ is the corresponding deviation, $\boldsymbol{V}_m^L$ is the eigenvector after modality m is converted to a fixed length, where $m \in \{image, text\}_{modality}$:

$$\boldsymbol{V}_m^L = \tanh(\boldsymbol{W}_{m1}\boldsymbol{V}_m + \boldsymbol{b}_{m1}) \tag{5}$$

The weighted score of the modality is shown in the following formula, where $\boldsymbol{W}_{fuse}$ represents the fusion weight, and $\boldsymbol{b}_{fuse}$ represents the fusion bias.

$$\boldsymbol{S}_{score} = \text{softmax}\left(\tanh\left(\boldsymbol{W}_{fuse}\left[\boldsymbol{V}_{image}^L, \boldsymbol{V}_{text}^L\right] + \boldsymbol{b}_{fuse}\right)\right) \tag{6}$$

The correlation between modality features is expressed as Eq. (7), where $\left(1 + \boldsymbol{S}_{image}\right)$ and $(1 + \boldsymbol{S}_{text})$ are the attention scores of the image modality and the text modality respectively, and $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ respectively denotes the fusion weight and fusion bias of the l-th layer of the fully connected layer, where $l = 8$:

$$r = \tanh\left(\boldsymbol{W}_l\left[\left(1+\boldsymbol{S}_{image}\right)\boldsymbol{V}^L_{image},\left(1+\boldsymbol{S}_{text}\right)\boldsymbol{V}^L_{text}\right]+\boldsymbol{b}_l\right) \tag{7}$$

The weight corresponding to the j-th vector of modality m is shown in Eq. (8), where j is the number of vectors contained in each modality, $\boldsymbol{V}^{L(i)}_m$ represents the j-th original vector of modality m, and $\boldsymbol{W}_{m2}$ and $\boldsymbol{W}_{m3}$ are the weights matrix, $\boldsymbol{b}_{m2}$ and $\boldsymbol{b}_{pm3}$ are the corresponding deviations.

$$\boldsymbol{\alpha}^{(j)}_m = \boldsymbol{W}_{m3}\tanh\left(\boldsymbol{W}_{m2}\boldsymbol{V}^{L(i)}_m+\boldsymbol{b}_{m2}\right)+\boldsymbol{b}_{m3} \tag{8}$$

Aggregating all the original vectors yields:

$$\boldsymbol{\alpha}_m = \text{softmax}\left(\boldsymbol{\alpha}_m\right) \tag{9}$$

The final output fixed-length one-dimensional fusion vector is:

$$\boldsymbol{V}_{fuse} = \sum\nolimits_m \boldsymbol{\alpha}_m \boldsymbol{V}^L_m \tag{10}$$

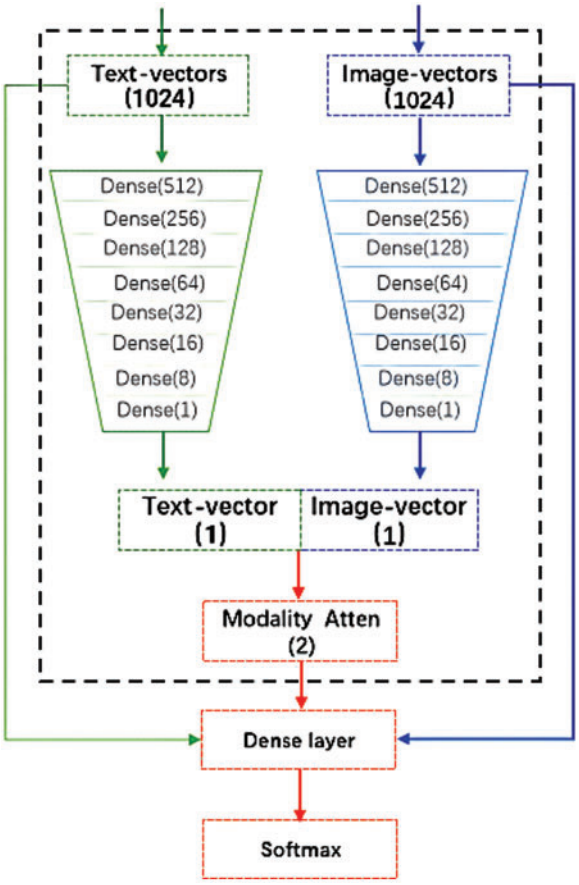

Figure 6: Schematic diagram of weight adaptive learning

4 Experiments

The experimental results are shown in Fig. 7. This section verifies the performance of the feature reconstruction output algorithm, and then introduces an experiment based on the weight adaptive

learning algorithm in the modality fusion. A comparative experiment of two types of algorithms highlights the performance of the algorithm proposed in this paper.

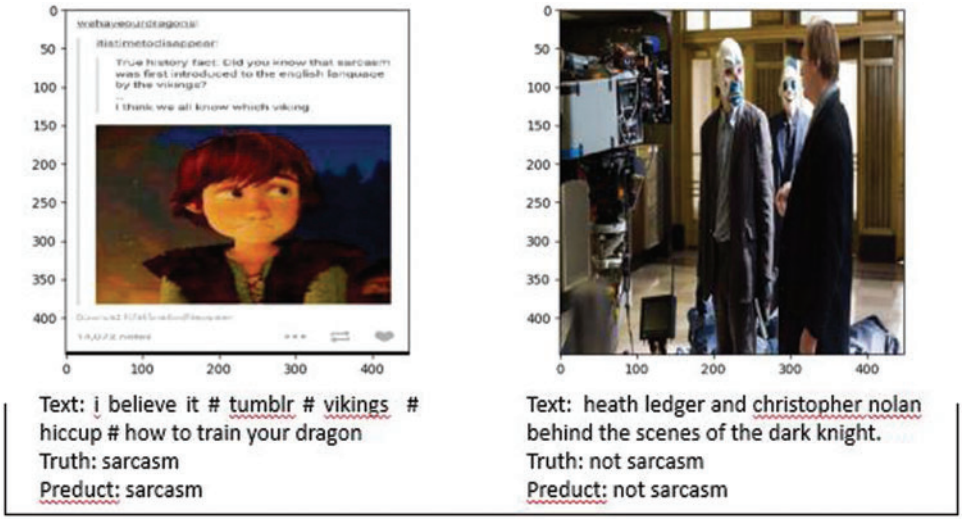

Figure 7: Schematic diagram of experimental effect

4.1 Research on Modal Feature Output Method Based on Sarcasm Detection Task

The objective of this section is to assess the efficacy of the feature reconstruction output algorithm. In order to facilitate the description of subsequent experiments, the data type, feature output model and modality category are used as the classification basis, as shown in Table 1. The fast dictionary learning (FDL) algorithm [25] will be used for detailed classification. The processing of images using singular value decomposition and a sparse dictionary allows for the display of the majority of the original image's feature information while simultaneously reducing the overall data volume. This reduction in data volume can effectively reduce the time spent on subsequent model training convergence.

Table 1: Experiment naming and classification

Classification basis	Classification details	Abbreviation
Data type	Raw data set without any algorithm processing	Raw data
	Raw data preprocessed by the FDL	FDL data
Feature output model	After the input data is learned by the model, the image feature vector and text feature vector are obtained, and these two modality vectors are directly output to the modality fusion part.	Feature direct output
	After learning the original image and original text feature vectors, feature reconstruction is performed to generate image vectors and text vectors containing cross-modality features	Feature reconstruction output

(Continued)

Table 1 (continued)

Classification basis	Classification details	Abbreviation
Modality class	Participate in experiments using only image modality data	Image modality
	Participate in experiments using only text modality data	Text modality

4.1.1 Benchmark Model

The text information in the input data is processed by the word embedding operation to obtain the original text vector, and the features in the text vector are learned through Bi-LSTM. The image modality in the input data is learned through ResNet to obtain image features. The two types of features are fused into a single feature vector, which is subsequently transmitted to the classifier in order to obtain the desired output. Since neural networks are unable to process natural language information directly, it is usually necessary to convert text data into vector form. This model uses Glove as a statistical model based on the global corpus to create the co-occurrence matrix. It constructs the co-occurrence matrix through the global corpus, uses the decay function to measure the distance between words, and uses the weighted least square method as the loss function. The model also considers the co-occurrence matrix information in the sliding window and the global word frequency statistics. By fully leveraging global information during training, the model is able to converge more efficiently and achieve superior outcomes in a shorter timeframe. The benchmark model's image processing component employs the ResNet architecture.

4.1.2 Dataset

Because there is a shortage of public datasets that closely match the research topic of this paper, we will use a dataset that contains bimodal data consisting of image and text from foreign Twitter platform public data. The experiment was carried out by Cai et al. [26]. The released satirical emotion-specific dataset, which is based on the user data of the Twitter platform, contains three modalities: image, text, and image-corresponding attributes. Combined with the research topic, two modality data of image and text are selected. This dataset is referred to as the Twitter image-text dual-modal dataset. The dataset contains 24,635 sets of data, including 19,816 sets of training set data, and 4819 sets of test set and verification set data. The specific distribution of the data is presented in Table 2:

Table 2: Data distribution of Twitter dataset

Training set	Test set	Validation set	Label (emotion)
8642	959	959	0 (Sarcastic)
11174	1450	1451	1 (Non-Sarcastic)
19816	2409	2410	Total

4.1.3 Data Preprocessing

Generally speaking, users can post comments on Twitter or other social platforms according to the way they are accustomed to speaking. As a result, these comments often contain colloquial language, which can cause errors in grammatical identification by machines. These errors can introduce

noise into the data when embedding text, ultimately affecting the vectorization of text information. Therefore, it is necessary to preprocess the data. For text data, the processing method is as follows:

1. Replace links and handles. The text information is often mixed with links that can jump to the third-party interface and the use of network symbols like "@username". This kind of data is meaningless to the research of this topic. Based on the principle of not missing data, use <url>, <user>replaced.
2. Delete punctuation, special symbols and stop words. This operation removes non-English words and some words that are not discriminative in the dataset.
3. Keep user tags. The text information published by some users contains "sarcasm", "irony", "humor" and other potentially ironic labels ("#tags"). Such artificial labels are very beneficial to model learning, so they need to be retained.

This paper utilizes the FDL algorithm for image data. By using a sparse dictionary to represent samples, the image preprocessed by FDL can basically display all the feature information of the original image, and its data volume is significantly reduced, which can effectively reduce the time taken for the subsequent model to converge.

4.1.4 Experimental Effect

A comparative experiment will be conducted on two types of modalities, image and text, to study the algorithm model with the best processing effect under certain fixed types of data.

The evaluation metrics employed in this study are precision and recall. In order to verify the performance of the model, comparative experiments will be carried out on the following four models:

1. Raw data simple output model: After inputting the raw data into the model to obtain the corresponding modality vector, it is directly output to the modality fusion module.
2. Original data reconstruction output model: After inputting the original data of the model to obtain the original vector, the feature reconstruction is performed first, and then output to the modality fusion module.
3. FDL simple output model: After the original data is learned by the FDL algorithm, the obtained FDL data is sent into the model to generate the corresponding modality vector, and the vector is directly output to the modality fusion module.
4. FDL reconstruction output model: After the original data is learned by the FDL algorithm, the obtained FDL data is sent to the model to generate the corresponding modality vector, and the vector is firstly reconstructed and finally output to the modality fusion module.

The experimental results of the FDL and feature reconstruction output algorithms proposed in this paper on the twitter single-mode data set are shown in Figs. 8 and 9, respectively. As can be seen from Fig. 8, when only the image modality data in the dataset is used, the performance of the FDL reconstruction output model is outstanding, with an accuracy rate of 87.6%, followed by the performance of the FDL simple output model, and the effect of other two types of models on the image dataset is close. It can be seen from Fig. 9 that when only the text modality data in the dataset is used, the FDL reconstruction output model and the original data reconstruction output model reach the optimum at the time of convergence, with an accuracy rate of 85.2%. Since FDL is only for image data preprocessing, it has no impact on text information. Therefore, it can be determined that the feature reconstruction output algorithm also plays an active role in the text modality data. In summary, it is concluded that the feature reconstruction output algorithm has good performance in this research.

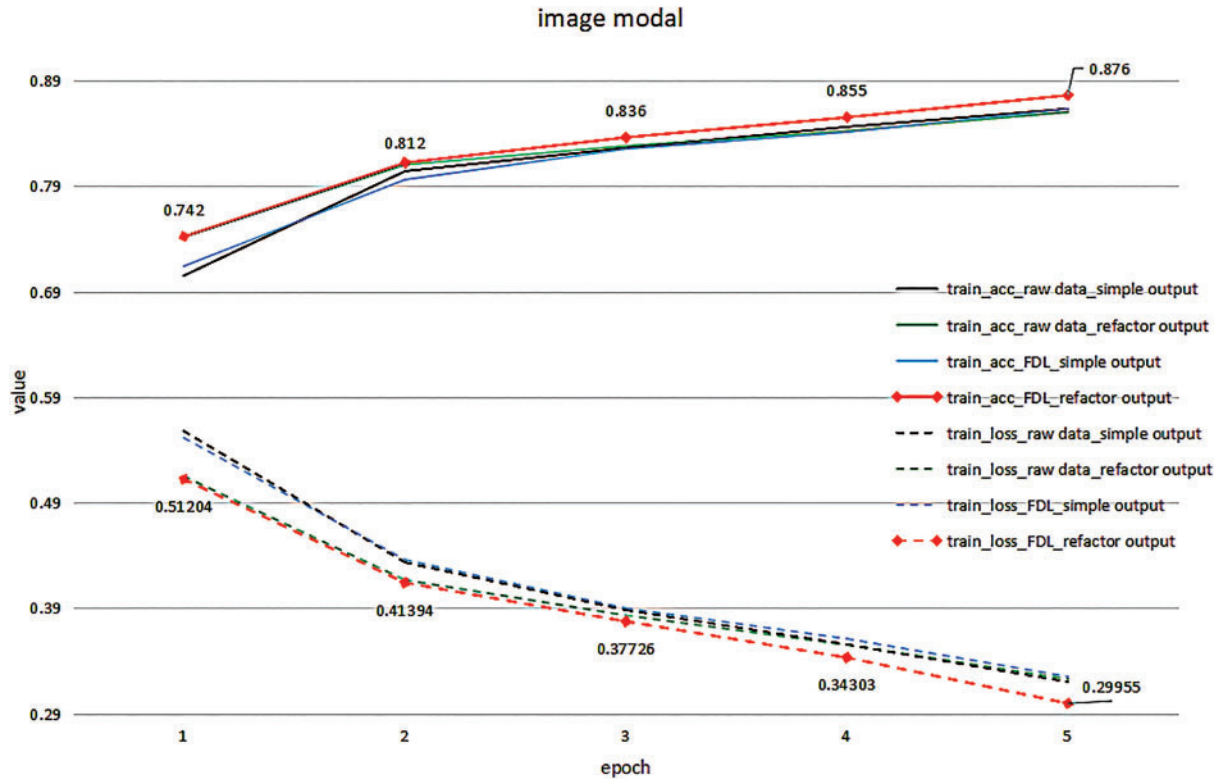

Figure 8: Image modality data comparison experiment

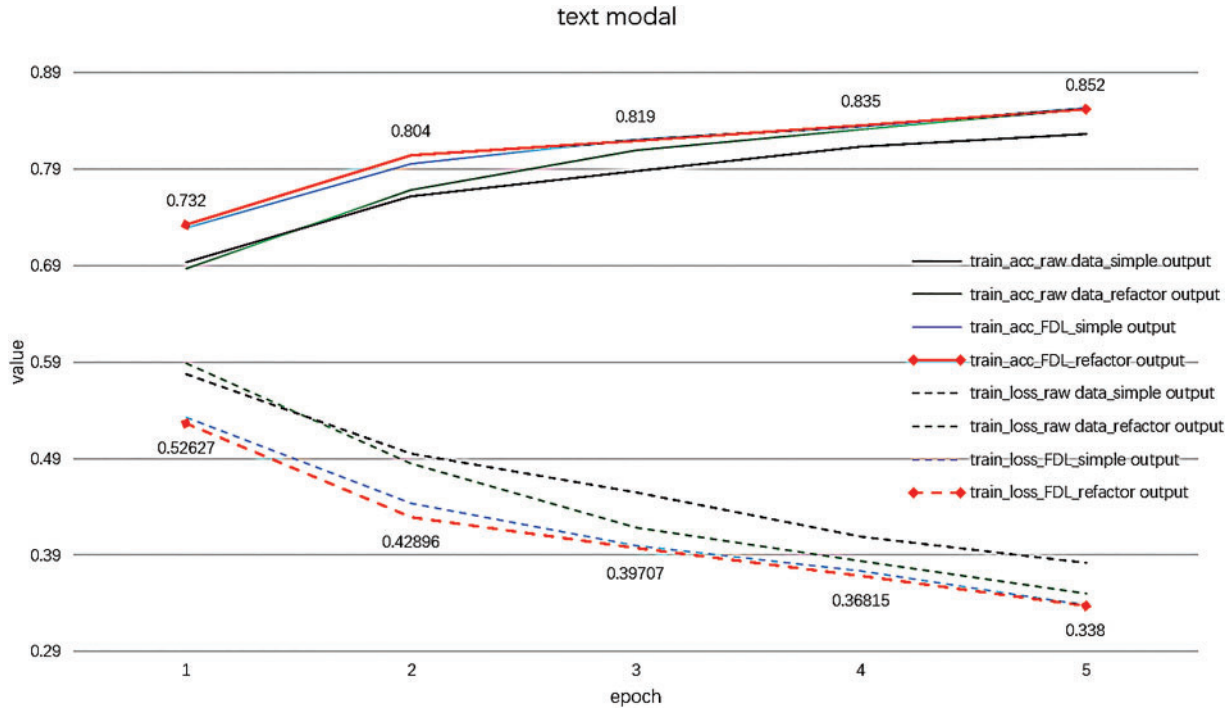

Figure 9: Text modality data comparison experiment

4.2 Research on Modality Fusion Method Based on Sarcasm Detection Task

The input of various classifiers is usually a one-dimensional vector. Therefore, it is crucial to develop an effective method for converting a multi-modal vector into a one-dimensional fusion vector output in order to achieve optimal performance in multi-modal classification tasks such as image-text fusion. This section will study the modality fusion method based on the sarcasm detection task. The weight ratio fusion method treats each modality separately, outputs them separately, and finally adds a certain weight to various outputs to form the final result. The second method is to construct a neural network to learn the weights of each component vector of the image modality, thereby obtaining a one-dimensional fusion vector. This type of method is called weight adaptive learning.

This section will verify the performance of the aforementioned two types of feature fusion algorithms. For the convenience of experimental description, the modality category is increased to four categories on the basis of Table 3, as shown in Table 3.

Table 3: Experiment naming and adding classification

Classification basis	Classification details	Abbreviation
Modality class	Participate in experiments using only image modality data	Image modality
	Participate in experiments using only text modality data	Text modality
	Simultaneously use two types of modalities of image and text, and the fusion method between modalities adopts weight adaptive learning algorithm	Weight learning modality
	Use both image and text modes at the same time, and the fusion method between the modalities is weight ratio fusion	Weight proportional fusion modality

This experiment focuses on whether the performance of the weight learning algorithm is excellent in the fusion comparison experiment of four types of data (original data through weight learning, FDL data through weight learning, original data through proportional fusion, and FDL data through proportional fusion) on the weight ratio fusion algorithm.

The experiment employs accuracy rate and loss value as evaluation metrics. Further comparative experiments will be conducted on four types of models (original data simple output model, original data reconstruction output model, FDL simple output model, FDL reconstruction output model).

In this experiment, the image and text feature weights are set to be distributed according to the ratio of 50% in the image and text weight ratio fusion method. In Figs. 10 and 11, 01 represents the original data simple output model, 02 represents the original data reconstruction output model, 03 represents the FDL data simple output model, and 04 represents the FDL data reconstruction output model. It can be seen from Fig. 10 that the model using weight adaptive learning has the highest accuracy rate among similar data. Under the refactoring, the weight learning effect achieves the best.

The multi-modal model is 0.3% more accurate than the best result achieved by the single image modality and 2.7% more accurate than the best result achieved by the text modality, confirming the effectiveness of multi-modal data for sarcasm detection.

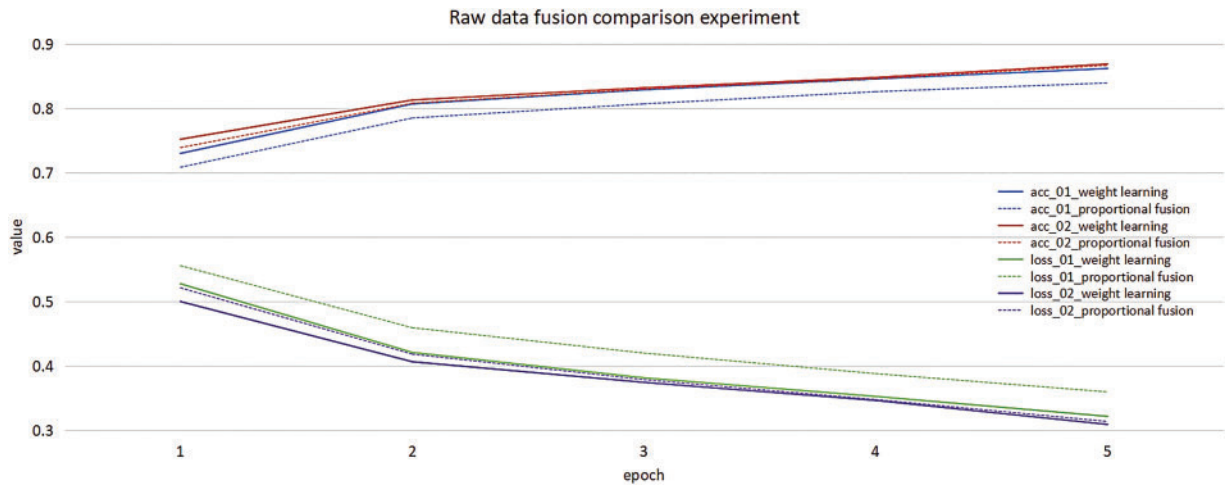

Figure 10: Raw data fusion comparison experiment

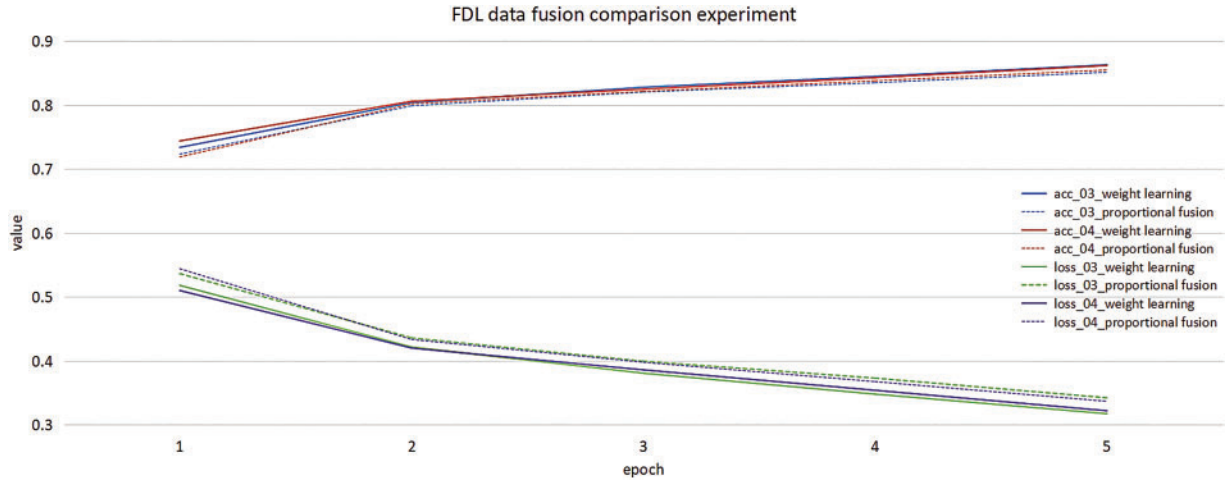

Figure 11: FDL data fusion comparison experiment

4.3 Comparison with Baseline Models

The accuracy of FDL data obtained by weight learning under feature reconstruction is optimal. The working performance of our model was evaluated in comparison with a series of state-of-the-art baselines:

1. Image-modality methods: **Image** [26] utilizes ResNet for sarcasm detection; **ViT**(Vision Transformer) [27] uses the [CLS] token representation as the input of pre-trained visual model to detect the sarcasm.
2. Text-modality methods: **TextCNN** [28] utilizes a CNN for text classification; **SIARN** (Self-Interactive Attention-based Recurrent Network) [29] uses inner-attention for text sarcasm detection; **SMSD** (Self-Matching Network for Sarcasm Detection) [30] designs a self-matching network to capture sarcasm incongruity information; **Bi-LSTM** [31] uses a bidirectional long-short memory network for text classification; **BERT** [32] is a pre-trained model which takes [CLS] text [SEP] as input for text classification.

3. Multimodel methods: **HFM** (Hierarchical Fusion Model) [26] proposes a multimodal hierarchical fusion model for sarcasm detection; **D&R Net** (Decomposition and Relation Network) [33] designs a decomposition and relation network by modeling cross-modality contrast and semantic association for sarcasm detection; **Res-BERT** (Residual BERT) [34] uses the BERT to encode text and combine text and image features for sarcasm prediction; **Att-BERT** (Attention BERT) [34] explores inter-modal attention and co-attention to model multimodal incongruity; **InCrossMGs** (Intra- and Inter-Modal Cross Media Graphs) [8] is a graph-based model and exploits both intra- and inter-modal sarcasm relations; **CMGCN** (Cross-Modal Graph Convolutional Network) [35] constructs a cross-modal graph convolutional network to draw the sarcasm reltions for sarcasm prediction.

As shown in Table 4, among all the methods, our method achieves the best and the most consistent working performance.

Table 4: Performance compared to state-of-the-art baselines with the evaluation metrics acc

Model		Acc (%)
Image	Image	64.76
	ViT	67.83
Text	TextCNN	80.03
	SIARN	80.57
	SMSD	80.9
	Bi-LSTM	81.9
	BERT	83.85
Multimodel	HFM	83.44
	D&R Net	84.02
	Res-BERT	84.80
	Att-BERT	86.05
	InCrossMGs	86.10
	CMGCN	86.54
	Ours	**87.9**

Despite the favorable outcomes of our experiments, there are instances where sarcasm is not accurately identified. The reasons for this discrepancy were analyzed in detail.

The detection of sarcasm is dependent on the context and the presence of subtle linguistic cues. If the aforementioned cues are not adequately captured or understood by the detection model, sarcasm may be misidentified. Furthermore, the prevalence of sarcasm is contingent upon the evolution of language trends and cultural references over time. Consequently, it is vital to update or train the detection model on a diverse range of data in order to detect more recent forms of sarcasm or sarcasm that is specific to certain cultures or communities. Fig. 12 illustrates an instance of sarcasm that was not identified due to a lack of external knowledge.

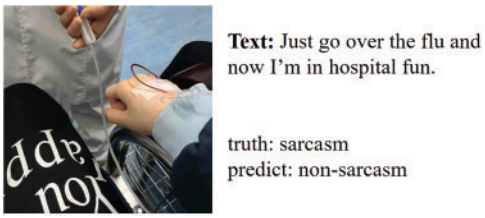

Figure 12: Examples of undetected sarcasm

With regard to the specific patterns of sarcastic tweets that are more readily identifiable than others, research has demonstrated that certain linguistic cues and patterns can be indicative of sarcasm. These include the use of irony, exaggeration, incongruity, and certain lexical and syntactic patterns.

5 Conclusion

The proposed algorithm for feature reconstruction in this paper, which is based on the attention mechanism, learns the low-rank features of another modality and calculates the weights of each original vector. It then obtains the reconstructed feature vector of the corresponding modality through weighted average. A comparison of methods shows that the feature reconstruction output algorithm has been effective in handling text modality data. Effective fusion of image data and text data in social media comment information can improve the performance of experiments. This paper proposes a weight adaptive learning algorithm that employs a neural network in conjunction with an attention mechanism to determine the relative importance of each modality, thereby facilitating weight-adaptive learning. The implementation of this feature fusion method has been shown to have a positive effect on the efficacy of sarcasm detection experiments. A comparison of the proposed method with baseline approaches revealed that it outperforms them.

Acknowledgement: None.

Funding Statement: This research was funded by National Key Research and Development Program of China (No. 2022YFC3302103).

Author Contributions: Study conception and design: Xiaofang Jin; data collection: Yinan Wu; analysis and interpretation of results: Yuying Yang; draft manuscript preparation: Ying Xu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data used in this research are available from the corresponding author, Y. Yang, upon reasonable request.

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

References

[1] GWI, "Social-GWI report," hubspotusercontent20.net, Feb. 23, 2022.

[2] R. Gonzálezibáñez, S. Muresan, and N. Wacholder, "Identifying sarcasm in Twitter: A closer look," in *Proc. Meet. Assoc. Comput. Linguist.: Human Lang. Technol.: Short Pap. Assoc. Comput. Linguist.*, Portland, Oregon, USA, Jun. 19, 2011.

[3] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, no. 6191, pp. 516–529, Sep. 1, 2017. doi: 10.1016/j.ins.2017.09.010.

[4] M. Zou, M. You, and T. Akashi, "Reconstruction of partially occluded facial image for classification," *IEEJ Trans. Electr. Electron. Eng.*, vol. 16, no. 4, pp. 600–608, Feb. 22, 2021. doi: 10.1002/tee.23335.

[5] Y. Sun *et al.*, "Discriminative local sparse representation by robust adaptive dictionary pair learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4303–4317, Jan. 14, 2020. doi: 10.1109/TNNLS.2019.2954545.

[6] X. Zhao, J. Huang, and H. Yang, "CANs: Coupled-attention networks for sarcasm detection on social media," in *Proc. 2021 Int. Joint Conf. Neural Netw. (IJCNN)*, Shenzhen, China, IEEE, Jul. 18–21, 2021, pp. 1–8.

[7] N. Ding, S. Tian, and L. Yu, "A multimodal fusion method for sarcasm detection based on late fusion," *Multimed. Tools Appl.*, vol. 81, no. 6, pp. 8597–8616, Mar. 1, 2022. doi: 10.1007/s11042-022-12122-9.

[8] B. Liang, C. Lou, X. Li, L. Gui, M. Yang and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proc. 29th ACM Int. Conf. Multimed.*, Dublin, Ireland, Oct. 17, 2021, pp. 4707–4715.

[9] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," arXiv preprint arXiv:2109.05184, Sep. 22, 2021.

[10] A. Shah and C. K. Maurya, "How effective is incongruity? Implications for code-mix sarcasm detection," arXiv preprint arXiv:2202.02702, Feb. 6, 2022.

[11] F. Ren *et al.*, "Hierarchical attention-based model with multi-task self-training for user profiling," in *Proc. 2021 IEEE Int. Conf. Big Data (Big Data)*, Orlando, FL, USA, IEEE, Dec. 15–18, 2021, pp. 500–509.

[12] G. Qiu, X. Yu, L. Jiang, and B. Ma, "Text-aware recommendation model based on multi-attention neural networks," in *Int. Conf. Knowl. Sci. Eng. Manag.*, Cham, Tokyo, Japan, Springer, Aug. 14–16, 2021, pp. 590–603.

[13] L. Lucas, D. Tomás, and J. Garcia-Rodriguez, "Exploiting the relationship between visual and textual features in social networks for image classification with zero-shot deep learning," in *Int. Workshop Soft Comput. Models Ind. Environ. Appl.*, Cham, Bilbao, Spain, Springer, Jul. 8, 2021, pp. 369–378.

[14] S. Gupta, A. Shah, M. Shah, L. Syiemlieh, and C. Maurya, "Filming multimodal sarcasm detection with attention," in *Int. Conf. Neural Inf. Process.*, Cham, Springer, Aug. 9, 2021, pp. 178–186.

[15] S. Kumar, A. Kulkarni, M. S. Akhtar, and T. Chakraborty, "When did you become so smart, oh wise one?! Sarcasm explanation in multi-modal multi-party dialogues," arXiv preprint arXiv:2203.06419, May 12, 2022.

[16] P. Kumar and G. Sarin, "WELMSD–word embedding and language model based sarcasm detection," *Online Inf. Review*, vol. 9, no. 7, pp. 1242–1256, Feb. 9, 2022. doi: 10.1108/OIR-03-2021-0184.

[17] E. Savini and C. Caragea, "Intermediate-task transfer learning with BERT for sarcasm detection," *Mathematics*, vol. 10, no. 5, pp. 844, Mar. 7, 2022. doi: 10.3390/math10050844.

[18] T. M. Tashu, S. Hajiyeva, and T. Horvath, "Multimodal emotion recognition from art using sequential co-attention," *J. Imaging*, vol. 7, no. 8, pp. 157, Aug. 21, 2021. doi: 10.3390/jimaging7080157.

[19] S. Pramanick, A. Roy, and V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, Hawaii, USA, Jan. 3–8, 2022, pp. 3930–3940.

[20] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, no. 3, pp. 279–294, Feb. 1, 2021. doi: 10.1016/j.future.2020.08.005.

[21] X. Yuan, L. Li, Y. A. W. Shardt, Y. Wang, and C. Yang, "Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4404–4414, Apr. 9, 2020. doi: 10.1109/TIE.2020.2984443.

[22] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3391–3402, Dec. 7, 2020. doi: 10.1109/TCSVT.2020.3043026.
[23] S. Miao *et al.*, "Balanced single-shot object detection using cross-context attention-guided network," *Pattern Recognit.*, vol. 122, no. 2, pp. 108258, Feb. 1, 2022. doi: 10.1016/j.patcog.2021.108258.
[24] M. H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, Mar. 15, 2022. doi: 10.1007/s41095-022-0271-y.
[25] X. Jin, Y. Wu, Y. Xu, and C. Sun, "Research on image sentiment analysis technology based on sparse representation," *CAAI Trans. Intell. Technol.*, vol. 7, no. 3, pp. 354–368, Jan. 4, 2022. doi: 10.1049/cit2.12074.
[26] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, Jul. 28–Aug 2, 2019, pp. 2506–2515.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[28] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 25–29, 2014, pp. 1746–1751.
[29] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with Sarcasm by Reading in-between," in *Proc. 56th Annu. Meet. Association for Computational Linguistics*, Melbourne, Australia, pp. 1010–1020, Jul. 15–20, 2018.
[30] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *World Wide Web Con.*, San Francisco, CA, USA, May 13–17, 2019, pp. 2115–2124.
[31] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," in *Neural Netw*, Montreal, QC, Canada, Jul. 31–Aug. 4, 2005.
[32] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deepbidirectional transformers for language understanding," in *Proc. 2019 Conf. N Am Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, Minneapolis, Minnesota, pp. 4171–4186, Jun. 2–7, 2019.
[33] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, Jul. 5–10, 2020, pp. 3777–3786.
[34] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Find. Assoc. Comput. Linguist.: EMNLP*, Nov. 19–20, 2020, pp. 1383–1392.
[35] B. Liang *et al.*, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.*, Dublin, Ireland, May 22–27, 2022, vol. 1, pp. 1767–1777.