Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078074

ARTICLE

Quantum-Inspired Complex-Valued Fusion Framework: Optimizing Intra-Modal Semantics and Inter-Modal Fusion in Multimodal Sarcasm Detection

Dong Zhang[1], Lianhe Shao[2,*], Weijie Xu[3], Xihan Wang[1,*] and Quanli Gao[2]

[1]School of Computer Science, Xi'an Polytechnic University, Xi'an, China
[2]School of Cybersecurity, Xi'an Polytechnic University, Xi'an, China
[3]State Grid (Xi'an) Environmental Protection Technology Center Co., Ltd., Xi'an, China

*Corresponding Authors: Lianhe Shao. Email: shaolianhe@xpu.edu.cn; Xihan Wang. Email: xihanwang@xpu.edu.cn

Received: 23 December 2025; Accepted: 26 March 2026; Published: 08 May 2026

ABSTRACT: With the popularization of multimodal content on social media, accurately identifying sarcastic intent is of great significance for understanding public attitudes and grasping public opinion trends. However, sarcastic expressions rely on context, exhibit inconsistencies in multimodal information, and have implicitly contradictory semantics. These characteristics pose challenges to traditional single-text modality methods. Existing multimodal methods, due to their default assumption of symmetric modal interactions and difficulty in capturing the subtlety of sarcasm and modal contradictions, yield limited detection performance. Therefore, this paper proposes a quantum-inspired complex-valued fusion framework to optimize the intra-modal semantics and inter-modal fusion in multimodal sarcasm detection. Firstly, this framework constructs a quantum-inspired complex-valued multimodal feature representation method. It embeds the text, visual, and audio modalities into the complex-valued Hilbert space, and models the feature intensity and directional information, respectively, through the two dimensions of "amplitude-phase", providing highly expressive basic features for fusion. Secondly, an asymmetric quantum interference fusion mechanism is designed. Based on the principle of quantum interference, a directional interference term and trainable parameters are introduced to accurately capture the asymmetric interaction relationship between modalities, where "text dominates semantic interpretation and vision supplements detailed evidence", effectively mining the modal contradictions on which sarcasm depends. Experimental results show that the F1-score of the proposed model has increased by 3.71% and 2.74% compared with M2Seq2Seq and SRLM, respectively, on the Mustard dataset. On the Memotion dataset, it also achieves performance improvements of 0.28% and 0.83% relative to M2Seq2Seq and SRLM. The effectiveness of the key modules in the model is also verified through ablation experiments.

KEYWORDS: Sarcasm detection; multimodal analysis; quantum interference

1 Introduction

In recent years, with the popularization of social media, users are increasingly inclined to convey their personal opinions and sentiment through multimodal content. In this context, it is crucial to accurately identify the satirical intention for gaining insight into the true attitude of the public towards hot issues and grasping the trend of public opinion. Sarcasm detection technology has shown broad application prospects in multiple fields. Despite the significant application value of sarcasm recognition, due to the fact that sarcastic expressions usually rely on context, the inconsistency among multimodal information, and the implicit semantic opposition, it has become a challenging problem in natural language processing and multimodal

analysis. Therefore, how to effectively integrate multimodal information to improve the accuracy of sarcasm recognition has become a key issue in current research.

Early sarcasm detection methods were mostly rule-based or based on a single text modality, and thus had difficulty fully capturing the linguistic complexity of sarcasm and cross-modal contradictions. With the popularization of multimodal data, researchers began to explore recognition methods that integrate text, visual, and speech information. For example, the M2Seq2Seq model proposed by Zhang et al. [1] models context dependence and multimodal fusion through intra modal and cross modal attention mechanisms, significantly improving the performance of sarcasm recognition. On the other hand, the quantum fuzzy neural network introduced by Tiwari et al. [2] attempts to utilize quantum circuits and fuzzy logic to enhance the ability to model the uncertainty of sarcastic features, further promoting the development of multimodal sarcasm detection.

Classical multimodal semantic interaction methods mostly assume a symmetric relationship, distributing modal contributions equally through attention. However, they overlook the dynamic dependence characteristic of "text-dominance and vision-supplementation" in sarcastic expressions. As a result, these methods have limitations in modeling the subtlety and inconsistency of sarcasm. Moreover, fusing heterogeneous modalities and capturing deep-context associations in real-world social media scenarios remain challenging. The advantages of quantum theory in uncertainty modeling and non-classical correlation capture offer new ideas for breaking through the limitations of classical machine learning. Therefore, this paper constructs a quantum-inspired multimodal sarcasm detection model, aiming to break through the bottlenecks of feature representation and interaction modeling in existing methods.

Furthermore, by constructing a quantum-inspired model that can effectively integrate multimodal information and accurately identify sarcastic intentions, this study aims to optimize upon existing research. On one hand, it utilizes quantum complex-valued embedding to optimize the intra-modal feature representation, depicting semantic intensity and contextual associations in the dual dimensions of "amplitude-phase". On the other hand, it improves the interaction-modeling strategy through the quantum interference mechanism to precisely capture the asymmetric dynamic dependencies between modalities. Ultimately, this approach enhances the accuracy and robustness of sarcasm detection.

1.1 Contribution

Propose a quantum-inspired complex-valued multimodal feature representation method. Embed text, visual, and audio modalities into the complex-valued Hilbert space. Co-model the feature intensity and directional information through amplitude and phase, strengthen the expression of sarcastic features, improve detection performance, and provide a highly expressive basic representation for subsequent fusion.

We design an asymmetric quantum interference fusion mechanism by using the quantum interference formula to model the directional interactions between different modalities. In view of the complex asymmetric interaction relationships among modalities, we introduce the principle of quantum interference into multimodal fusion. Through the directional interference terms and trainable parameters, we can accurately capture the contradictions and inconsistencies between modalities upon which sarcasm depends.

1.2 Organization

The organization of this article is as follows: Section 2 reviews the relevant literature and work comprehensively, Section 3 introduces the methods used in this study in detail. Section 4 discusses and analyzes the experimental results. In Section 5, we draw conclusions based on the experimental results and summarize them.

2 Related Work

2.1 Sarcasm Detection

Sarcasm detection is the task of identifying sarcastic expressions in natural languages—a focus of research due to sarcasm's tendency to conceal true intentions, with applications in human-machine interaction and social media analysis [3]. Early approaches relied on rule-based methods (e.g., [4,5]): these depended on manually crafted features, incurring high algorithmic costs and limited adaptability. The field shifted to multimodal solutions in 2016, when Schifanella et al. [6] proposed two visual-textual fusion frameworks, marking the first cross-modal attempt for sarcasm detection. By 2020, multimodal methods expanded: Srivastava et al. [7] designed a deep learning-based hierarchical context method to mine dialogue history layer by layer; Wang et al. [8] proposed a vision-language network (using BERT and ResNet for text-graphic feature extraction) for sarcasm detection.

Subsequent advances focused on three core directions. First, multimodal fusion & semantic enhancement: Cai et al. [9] (2019) used hierarchical fusion to boost model expressiveness; Wu et al. [10] (2021) introduced a text-centered TCSP framework to enhance discourse semantics; Ding et al. [11] (2022) proposed a three-level late-fusion model (with residual connections) to unify text/audio/video features while preserving original and deep semantics. Second, incongruity modeling (sarcasm's core cue): Pan et al. [12] (2020) captured cross-modal inconsistencies via inter-modal attention; Lu et al. designed a dual multimodal contrastive attention model to learn sarcastic word scores and inter-modal inconsistencies; Liang et al. [13] (2022) used graph convolutions for contextual dependencies; Qiao et al. [14] solved the problem of irrelevant information interference and important information omission in multimodal sarcasm detection through the inconsistency learning module guided by local semantics, the global inconsistency learning module and their mutual reinforcement module. Wen et al. [15] (2023) modeled inconsistencies from semantic-emotional perspectives; Zhang et al. [3] (2024) adopted a gating network for multimodal contextual and task knowledge inconsistencies. Third, practical optimizations: Bedi et al. [16] (2021) built a Hindi-English dataset and used attention-enriched architectures for utterance representation; Dubey et al. [17] (2025) fused BERT-generated conversation summaries with metadata; Lee et al. [18] (2020) and Gao et al. [19] (2024) proposed sample augmentation techniques (e.g., contextual generation, back-translation).

2.2 Quantum Inspired Machine Learning

Quantum Inspired Machine Learning (QML) integrates quantum theory's mathematical tools into classical machine learning/deep learning frameworks, enabling modeling of non-classical, human-cognitive task characteristics on classical hardware—it retains deep networks' learning ability while aligning with human cognition to address core task challenges. The field's foundation was laid in 1995 (Kak [20]'s quantum neural computing concept), spawning algorithms like quantum neural networks.

For multimodal sarcasm detection, QML addresses uncertainty and cross-modal correlation challenges: Zhang et al. [21] (2021) proposed a quantum-inspired complex fuzzy network to handle sarcastic language ambiguity; Liu et al. [22] (2021) used complex-valued representations and quantum interference to model cognitive uncertainty; Tiwari et al. [2] (2024) enhanced multimodal sarcasm feature expression via a quantum-fuzzy combination. Yan et al. [23] (2026) proposed Lindblad-QNN and Liouville-QNN, which model multimodal data as Markovian/non-Markovian open quantum systems via Lindblad Master Equation (LME) and Stochastic Liouville–von Neumann Equation (SLNE) for sentiment analysis and sarcasm detection. Yan et al. [24] (2025) developed LLQNN, a quantum-inspired neural network integrating LME and complex-valued LSTM, to model text-image multimodal interactions for sentiment analysis and sarcasm detection, enhancing feature fusion effectiveness and model interpretability via von-Neumann entanglement entropy.

Recent quantum-inspired multimodal works rely on simulating quantum system dynamics to achieve cross-modal fusion, with a focus on modeling the interaction between quantum systems and the environment. Our work, by contrast, draws inspiration from the mathematical form of the quantum interference formula, constructing complex-valued embeddings via amplitude-phase representation and designing asymmetric fusion through complex-valued calculations, which directly captures modal interactions without involving complex quantum physical modeling. This design focuses on the essential demand of multimodal sarcasm detection for implicit correlation mining, providing a simple and effective quantum-inspired solution.

3 Methodology

The overall framework includes encoder and decoder, as shown in Fig. 1.

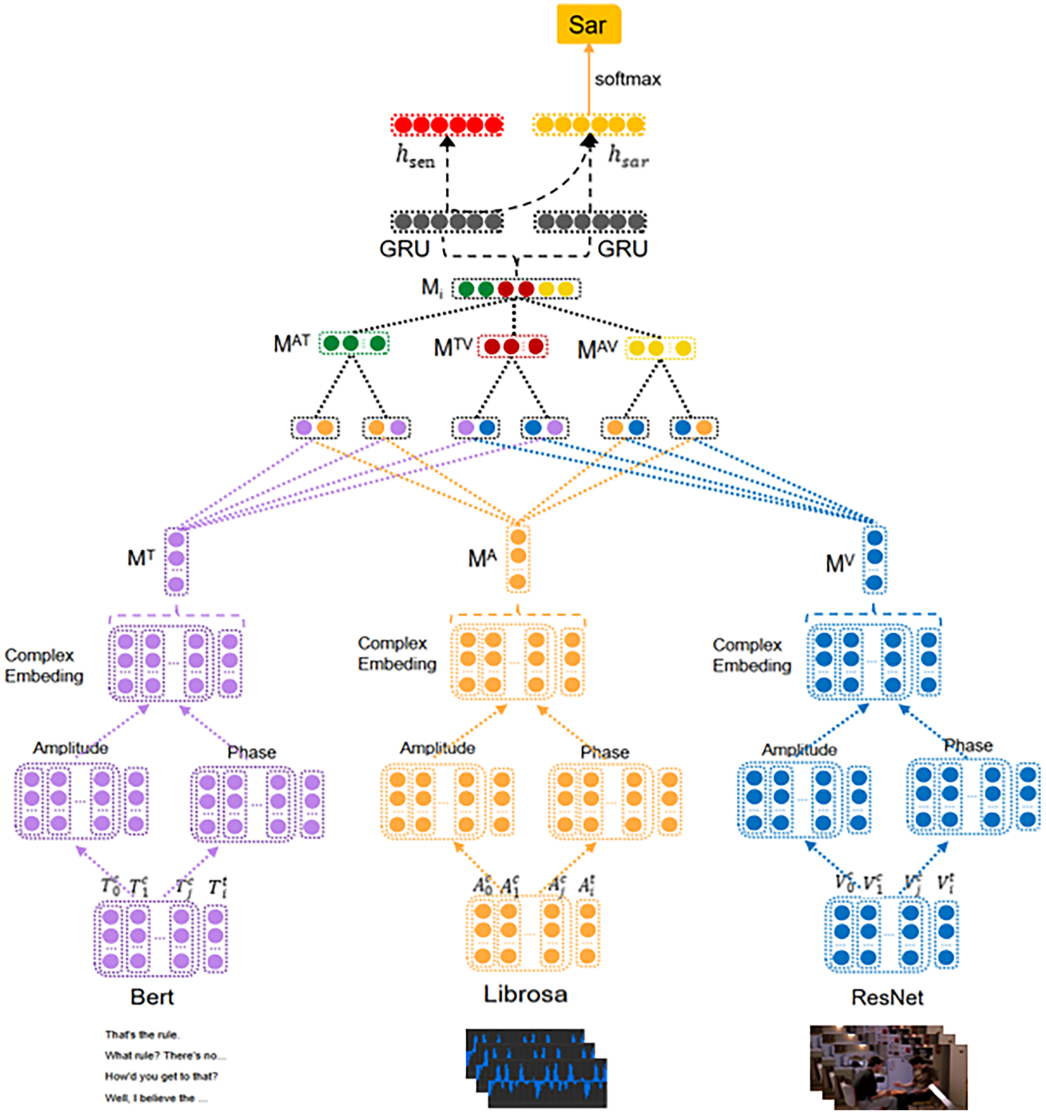

Figure 1: Overall structure diagram.

3.1 Problem Description

Through feature extraction, the original information from the three modes is divided into target features and context features. The target feature of text is set as T_i^t, the context feature is set as T_j^c, the target feature of image is set as V_i^t, the context feature is set as V_j^c, the target feature of audio is set as A_i^t, and the context feature is set as A_j^c. Here, $i \in \{0, 1, \ldots, N-1\}$, N represents the number of samples; $j \in \{0, 1, \ldots, n-1\}$, n represents the number of contexts. The goal of this algorithm is to predict sarcasm labels and detect sarcasm.

3.2 Encoder

The main function of the encoder is to aggregate the context features of a single modality using intra attention and fuse the features of different modalities using inter modality attention.

3.2.1 Complex Value Text, Visual and Audio Embedding

Modeling the information of text, vision and audio modes as superposition of complex embeddings. Suppose that the Hilbert space Ht of a text consists of a set of orthogonal ground states $\{|w_t^j\rangle\}_{j=1}^n$, with word as the basic semantic unit, the jth word w_t^j can be expressed by the base state $|w_t^j\rangle$. The set of ground states of this group of basic words is $\{|w_t^1\rangle, |w_t^2\rangle, \ldots, |w_t^n\rangle\}$, ξ^{th} target statement u_t^ξ is modeled as a complex superposition embedding of these ground states, expressed as:

$$|u_t^\xi\rangle = \sum_{j=1}^{n} z_t^j |w_t^j\rangle, z_t^j = r_t^j e^{i\theta_t^j} \tag{1}$$

where n is the number of words in the discourse, z_t^j is the amplitude of complex weight coefficient in polar coordinates, r_t^j is the module of complex value, called amplitude, and θ_t^j is the phase of z_t^j.

When constructing the complex vector of the ξ^{th} utterance, first extract the text features as amplitude, then process the basic amplitude features through the linear layer, GELU activation function, layer normalization and Dropout, and after the transformation of the hidden layer, generate the phase $\phi \in [-\pi, \pi]$ through the linear layer and the Tanh activation function. Based on Euler formula $e^{i\phi} = \cos\phi + i\sin\phi$, the real part is composed of the product of amplitude and phase cosine ($real = r \cdot \cos\phi$). The imag part is composed of the product of amplitude and phase sine ($imag = r \cdot \sin\phi$). The combination of the two forms a complete complex vector $|u_{t\xi}\rangle = r \cdot \cos\phi + i \cdot r \cdot \sin\phi$, in order to capture the long-term association of semantic units in complex valued space, complex valued attention mechanism is further introduced. The real part and imag part are used as query, key and value for cross interaction. The real part query matches the imag part key value, and the imag part query matches the real part key value. The important association is strengthened through attention calculation. The calculation process is as follows:

$$A_{real} = \text{Attention}(real, imag, imag) \tag{2}$$

$$A_{imag} = \text{Attention}(imag, real, real) \tag{3}$$

$$f_t^\xi = A_{real} + i \cdot A_{imag} \tag{4}$$

The final complex valued text feature f_t^ξ is obtained through Formula (4), providing a more expressive text representation for subsequent multimodal fusion.

For visual and audio features, a complex feature construction process similar to the text mode can be used to construct duplicate vectors through amplitude phase, and the final complex feature can be obtained through attention. Image features are represented as f_i^ξ, and audio features are represented as f_a^ξ.

3.2.2 Intra Modal Feature Fusion

Intra modal attention is used to aggregate contextual features in a single mode and fuse them with target features. Different context features have different effects on target features. Attention in the mode can highlight context features with higher contribution through adaptive weight allocation.

Let $U \in \{T, V, A\}$, U_i^t represents the target feature, U_j^c represents the context feature, and amplitude reflects the strength of the complex value feature. With it, you can measure the degree of correlation between the context feature and the target feature, and then assign attention weight to the context feature to achieve adaptive aggregation of context information that has more contributions to the target.

$|U_i^t|$ represents the amplitude of the target feature, $|U_j^c|$ represents the amplitude of the context feature. Use the attention mechanism to calculate the weight of each context feature to the target feature. The calculation process is as follows:

$$E_{i,j} = \exp\left(\tanh\left(|U_i^t| + |U_j^c|\right) + 10^{-7}\right) \tag{5}$$

$$\alpha_j = \frac{E_{i,j}}{\sum_{k=1}^{m} E_{i,k}} \tag{6}$$

$$C_{real} = \sum_{j=1}^{m} \alpha_j \cdot U_{c,real}^j \tag{7}$$

$$C_{imag} = \sum_{j=1}^{m} \alpha_j \cdot U_{c,imag}^j \tag{8}$$

$$C = C_{real} + i \cdot C_{imag} \tag{9}$$

The weight of each context feature to the target feature can be calculated through Formulas (5) and (6). $E_{i,j}$ represents the attention score of the jth context feature relative to the ith target feature, α_j represents the normalized weight of the jth context feature to the target feature, where m represents the total number of context features, C_{real} and C_{imag} represent the weighted aggregation results of the real and imag parts, respectively, and C represents the final aggregated complex value context feature.

M^U is used to represent the aggregation feature of context feature and target feature, W_s and W_p are trainable weights, and then residual connection is used to fuse C and target feature U_i^t.

$$M^U = U_i^t + \tanh\left(W_s c_a + W_p U_i^t\right) \tag{10}$$

The residual connection in Formula (10) is used to alleviate the problem of gradient disappearance, improve the training effect, and enhance the efficiency of representation learning.

3.2.3 Inter Modal Feature Fusion

Cross modal attention mechanism is used to fuse features of different modes. Various modes interact with each other, and cross modal attention can find the degree of influence between different modes. The speaker's subjective attitude can be seen as the joint expression of the two quantum superposition states of the three modal features of text (T), vision (V) and audio (A), as shown in Fig. 2. Taking text vision as an example, the formula can be expressed as:

$$\varphi_{tv}(x) = \alpha\varphi_t(x) + \beta\varphi_v(x) \tag{11}$$

where $\varphi_t(x) = M^t$ and $\varphi_v(x) = M^v$ (M^t, M^v represent the results after modal internal fusion) represent the complex probability amplitude of text and visual representation, respectively, $f_t(x) = |\alpha\varphi_t(x)|^2$ and $f_v(x) = |\beta\varphi_v(x)|^2$ represent the corresponding probability distribution, and the joint probability distribution after text visual fusion can be expressed as:

$$\begin{aligned} P(x) &= |\varphi_{tv}(x)|^2 = |\alpha\varphi_t(x) + \beta\varphi_v(x)|^2 \\ &= |\alpha\varphi_t(x)|^2 + |\beta\varphi_v(x)|^2 \\ &+ 2|\alpha\beta\varphi_t(x)\varphi_v(x)|\cos\Delta\phi \end{aligned} \tag{12}$$

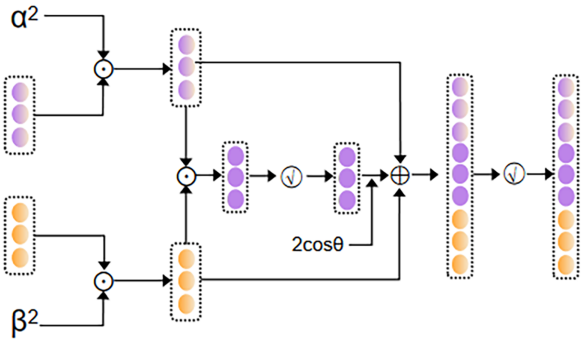

Figure 2: Quantum interference fusion.

To adapt to the asymmetric interaction characteristics between the text and visual modalities, where "text semantics constrain visual interpretation and visual information supplements text ambiguity", we design the text-visual feature fusion into two directional interference fusion patterns: One is text-dominant fusion. Based on the complex-valued probability amplitude $\varphi_t(x)$ of text features, the phase-difference calculation is $\Delta\phi_{tv} = \theta_t(x) - \theta_v(x) + \text{param}$ ($\theta_t(x)$ represents the phase information of the text, $\theta_v(x)$ represents the phase information of the image). This focuses on depicting the semantic constraint relationship of text on vision. The other is vision-dominant fusion. Based on the complex-valued probability amplitude $\varphi_v(x)$ of visual features, the phase-difference calculation is switched to $\Delta\phi_{vt} = \theta_v(x) - \theta_t(x) + \text{param}'$. This focuses on capturing the supplementary effect of vision on text ambiguity. The interference term is implemented by first extracting the amplitude (via torch.abs() on complex embeddings) and phase (via torch.angle()) from M^t and M^v, then adding the trainable phase biases θ/θ' to the basic inter-modal phase difference to obtain the final phase difference, and finally calculating the cross-modal interactive correlation via the product of weighted amplitudes and the cosine of the final phase difference. The interference term is further combined with the squared weighted amplitudes of the two modalities to generate the fused feature. Here, both param and param' are trainable parameters, used to quantify the semantic constraint intensity in text-dominant cases and the ambiguity-supplement degree in vision-dominant cases, respectively. Meanwhile, α and β, trainable parameters with initial values set to $\alpha^2 = 0.7$ and $\beta^2 = 0.3$ (meeting the initial constraint of $\alpha^2 + \beta^2 = 1$) responsible for weighting the overall contributions of the two modalities in the interference fusion. Precise modeling of modal interaction characteristics is achieved through their asymmetry.

Following the asymmetric design mentioned above, we can divide the feature fusion of the three modes into text-vision, vision-text, text-audio, audio-text, audio-visual, visual-audio, as shown in Fig. 3.

$$M_i = concat(M^{TV}, M^{VT}, M^{AV}, M^{VA}, M^{TA}, M^{AT}) \tag{13}$$

The fusion feature M_i of three modes is obtained from Formula (13). Considering the redundancy of multimodal feature splicing, the drop out layer is introduced here for regularization to avoid over fitting. The processed Mi will continue the subsequent process as the input of the decoder.

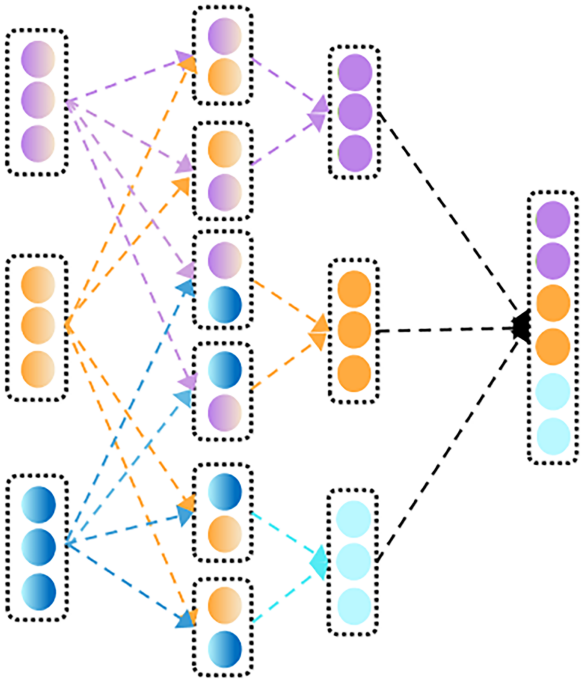

Figure 3: Inter modal feature fusion.

3.3 Decoder

The decoder function decodes the fused features into sarcasm and sentiment features. Self-attention mechanism and Gated Recurrent Unit (GRU) are used to distinguish these two characteristics. GRU is a recursive neural network (RNN), which is proposed to solve the problems of long-term memory and gradient in back-propagation. GRU is used to mine hidden features in fusion features.

$$I_{sen} = GRU(M_i) \tag{14}$$

$$I_{sar} = GRU(M_i) \tag{15}$$

In Eqs. (14) and (15), hidden sarcasm and hidden sentiment features are represented as I_{sen} and I_{sar}. Emotional features contribute to the detection of Sarcasm features. Therefore, some hidden emotional features are integrated into satirical features. For emotional features, self attention mechanism is used to extract important parts from features:

$$E_{sen} = \tanh(W[I_{sen}; I_{sen}]) \tag{16}$$

$$\alpha_{sen} = \text{Softmax}(w^T E_{sen}) \tag{17}$$

$$h_{sen} = \alpha_{sen} I_{sen} \tag{18}$$

The above three equations represent the process of self attention mechanism. In Eq. (16), E_{sen} is the attention score. In Eq. (17), α_{sen} is the attention weight. In Eq. (18), h_{sen} is a weighted representation of sentiment features. Let t_{sen} be the hidden feature of M_i in Formula (14) GRU. Therefore, we have:

$$I_c = \text{concat}(I_{sar}, t_{sen}) \tag{19}$$

$$E_{sar} = \tanh(W[I_c; I_c]) \tag{20}$$

$$\alpha_{sar} = \text{Softmax}(w^T E_{sar}) \tag{21}$$

$$h_{sar} = \alpha_{sar} I_c \tag{22}$$

In Eq. (19), first connect the sarcasm feature and the implicit sentiment feature to get I_c. Use the hidden feature in the sentiment feature to increase the expressiveness of the Sarcasm feature. (20)–(22) apply the self attention mechanism to I_c to obtain the weighted satirical feature h_{sar}.

3.4 Implementation and Training Details

Specific configurations for the feature extraction phase are as follows: For text features, the bert-base-uncased pre-trained model variant is adopted, with its weights kept frozen throughout the process (no fine-tuning is performed); the maximum input sequence length is set to 128, and preprocessing includes tokenization, length truncation, and zero-padding. Finally, average pooling is applied to the outputs of the last 4 Transformer layers of the [CLS] token in each utterance, yielding a 768-dimensional text feature representation. Visual features are extracted using the ResNet-152 model pre-trained on the ImageNet-1k dataset. The video sampling strategy involves dividing each video into 3 non-overlapping segments of equal duration and extracting the first frame of each segment; frame preprocessing includes resizing to 224 × 224 resolution and applying ImageNet normalization. After removing the final fully connected layer of the model, a 2048-dimensional visual feature per frame is obtained. Audio features are extracted as Mel spectrograms using the Librosa library, with parameters: sampling rate of 16 kHz, n_mels = 128, fmax = 8000 Hz, window length = 25 ms, and hop length = 10 ms. Preprocessing includes removing vocal interference via short-time Fourier transform (STFT) combined with soft mask separation; after adjustment via zero-padding or uniform sampling, a 128-dimensional audio feature is obtained.

The construction details of the complex-valued representation are as follows: Text features extracted from frozen bert-base-uncased (768D), visual features from ImageNet-pretrained ResNet-152 (2048D), and audio features (128D) are sequentially chunked and projected to a unified dimension of 128, yielding multi-modal sequential features with the same shape [B, L, 128]. The amplitude layer is derived from the temporal output of the Bi-GRU, while the phase vector is generated by a phase prediction network (comprising 2 fully connected layers and ReLU activation functions), which maps the amplitude feature to the range $[-\pi, \pi]$. The phase vector is then converted into cosine and sine coefficients via trigonometric mapping, which are element-wise multiplied with the amplitude feature to serve as the real and imaginary parts of the complex-valued representation—this achieves phase-amplitude alignment. The complex-valued attention uses a single-head mechanism, following the scaled dot-product formula: The query (Q), key (K), and value (V) are constructed by cross-using the real and imaginary parts of the complex-valued representation.

This study focuses on multimodal sarcasm detection, using a training framework with an auxiliary sentiment analysis signal, where loss weights are assigned as 0.7 for the core sarcasm detection task and 0.3 for the auxiliary sentiment loss to guide optimization. The training batch size is set to 256, with a maximum of 60 epochs. An early stopping mechanism is employed using the validation set F1-score as the core metric—training terminates if the improvement is ≤0.001 for 5 consecutive epochs, and the counter resets when validation loss decreases. To ensure result reliability and reproducibility, all random operations use four fixed

seeds (42, 88, 131, 230). The Adam optimizer is adopted with a weight decay of 0.05, and the learning rate follows a strategy of linear warm-up for the first 5 epochs, followed by a 20% decay every 10 epochs.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

The Mustard and Memotion datasets are used for the evaluation of the proposed model. Both datasets are multimodal and have sarcasm and sentiment labels. The specific sizes and labels of the datasets are shown in Table 1. The Mustard dataset consists of 690 data entries, which include video clips of character dialogues from TV series such as 'The Big Bang Theory' and 'Friends', along with corresponding text information. It has 2 sarcasm labels and 3 sentiment labels. The training set is divided as follows. The Memotion dataset contains 6992 data entries composed of text and social media-sourced images. Since no effective context exists for this dataset, we directly disable the intra-modal context module in our model. The images are sourced from social media. The number of sarcasm labels in Memotion is 2, and the number of sentiment labels is 3. To ensure the fairness of model evaluation, the test set of Memotion is constructed as a balanced test set using stratified under sampling, avoiding the interference of the original imbalanced distribution on performance assessment. The division of the datasets is presented in Table 2:

Table 1: Statistics of data sets.

Dataset	Task	Label	NOS	Proportion (%)
Mustard	Sarcasm	True	345	50
		False	345	50
	Sentiment	Positive	210	30.4
		Neutral	89	12.9
		Negative	391	56.7
Memotion	Sarcasm	True	5448	77.9
		False	1544	22.1
	Sentiment	Positive	631	9.0
		Neutral	2201	31.5
		Negative	4160	59.5

Table 2: The division of the datasets.

Dataset	Task	Label	NOS
Mustard	Train	True	248
		False	248
	Val	True	48
		False	49
	Test	True	49
		False	48

(Continued)

Table 2 (continued)

Dataset	Task	Label	NOS
Memotion	Train	True	4830
		False	926
	Val	True	309
		False	309
	Test	True	309
		False	309

4.1.2 Hyperparameter Settings

Hyper parameter settings in this article are shown in the Table 3:

Table 3: Hyperparameter settings.

Hyper parameter	Mustard	Memotion
Text features (dims)	768	768
Image features (dims)	2048	2048
Audio features (dims)	128	/
Learning rate	0.0005	0.0015
Dropout	0.3	0.3
Activation Func	Softmax	Softmax

4.1.3 Evaluation Metrics

We use four indicators to evaluate the performance of sarcasm detection: Precision, recall, F1-score, and acc to evaluate model performance. In these four indicators, F1-score is usually used to deal with binary classification problems. Therefore, we choose F1-score as our main evaluation index and others as our secondary evaluation index.

4.1.4 Baseline Methods

UPB-MTL [25]: It is a multimodal and multitask learning architecture, which combines Albert for text coding and VGG-16 for image representation

RCNN-RoBERTa [26]: It uses pre trained RoBERTa vector and RCNN to capture context information.

QFNN [2]: It is a multimodal multi task learning architecture, which combines classical and quantum neural networks (QNN) and Fuzzy logic.

A-MTL [27]: It proposes a multi task model based on attention, which can simultaneously analyze emotion, sentiment and detect sarcasm.

M2Seq2Seq [1]: M2Seq2Seq encoder is designed with double tension mechanism. It can learn the dynamics within and between modes at the same time, so as to realize multi tag classification including sentiment, emotion, and sarcasm.

SRLM [3]: The single input stream adaptive representation learning model realizes adaptive representation learning based on a gating network. It fuses serialized image features and text features,

configures dedicated expert networks for sarcasm and sentiment classification, and acquires weighted shared knowledge.

QPM [22]: The multitask learning framework based on quantum probability (QP) combines quantum superposition states, quantum interference and quantum incompatibility measurement for multimodal representation of each discourse.

4.2 Performance Analysis

4.2.1 Comparison of Different Models

As shown in Table 4, the performance of our model in different environments on Mustard and Memotion datasets has reached the level of classic networks. In Mustard dataset, our model F1 value reached 78.31 ± 0.83%, 3.71% higher than M2Seq2Seq, 2.74% higher than SRLM. In the Memotion dataset, the performance of the model is not as good as that of the Mustard dataset. F1 reaches 62.15 ± 0.13%, 0.28% higher than M2Seq2Seq, and 0.83% higher than SRLM, acc reaches 62.13 ± 0.13%. And under the standard division, F1 is 51.77 ± 0.24%, F1 (weighted) is 56.33 ± 0.38%, acc is 52.28 ± 0.41%. These performance advantages are mainly attributed to: the complex value embedding of multimodal features retains more abundant intra modal feature expressions, the quantum interference fusion between modes captures cross modal implicit correlation, and the asymmetric fusion strategy designed for different modal interaction characteristics further optimizes the cross modal information collaboration effect.

Table 4: Comparison of performance of different models.

Mustard				Memotion			
Model	**Pre**	**Recall**	**F1**	**Model**	**Pre**	**Recall**	**F1**
UPB-MTL	65.12	65.41	65.41	UPB-MTL	51.38	51.71	51.59
RCNN-RoBERTa	68.70	64.33	65.16	RCNN-RoBERTa	50.44	50.77	50.52
QFNN	68.87	68.88	68.88	QFNN	51.77	51.67	51.88
A-MTL	73.40	72.75	72.57	A-MTL	60.23	59.74	59.85
M2Seq2Seq	74.51	74.69	74.60	M2Seq2Seq	61.94	61.68	61.87
SRLM	75.46	75.69	75.57	SRLM	61.21	61.44	61.32
Transformer_fusion	72.24	72.13	72.18	Transformer_fusion	58.74	58.73	58.74
QPM	77.41	77.61	77.53	QPM	61.42	61.07	61.39
Ours	**78.51 ± 0.79**	**78.32 ± 0.87**	**78.31 ± 0.83**	**Ours**	**62.17 ± 0.14**	**62.13 ± 0.13**	**62.15 ± 0.13**

4.2.2 Effects of Different Data Modes

It can be seen from the data in Table 5 that when the information of text, audio and visual modes is used at the same time, the performance of sarcasm detection reaches the optimal level, with F1 score of 78.31 ± 0.83%, F1 (weighted) is 78.31 ± 0.86% and accuracy rate of 78.46 ± 0.70%, which fully demonstrates the significant value of multimodal information collaborative fusion in improving the effect of sarcasm detection. In the comparison of dual-mode combination, the fusion performance of text and image is the most prominent. The F1 score reaches 72.14 ± 0.58%, F1 (weighted) is 71.76 ± 0.49% and the accuracy rate is 71.90 ± 0.51%, which is better than the fusion effect of audio and image, text and audio. This indicates that the semantic information of text and the visual information of image have strong complementarity in the satirical detection task, and can more effectively capture the characteristics of satirical expression. In the

single mode detection, the performance of text information is far more than that of images and audio, with an F1 score of 69.59 ± 0.54%, F1 (weighted) is 69.59 ± 0.69% and an accuracy rate of 69.54 ± 0.55%, which means that ironic factors such as ironic words, special sentence structures and other ironic factors in the text are easier to be identified, and are the core information source of ironic detection; The image mode may ignore the subtle changes of human expression due to the frame extraction method, resulting in the loss of information, and the performance is limited to a certain extent. The audio mode has the worst performance due to the low recognition of intonation features.

Table 5: Effects of different data modes.

	Mustard		
Model	**F1**	**F1 (weighted)**	**ACC**
T + A + V	78.31 ± 0.83	78.31 ± 0.86	78.46 ± 0.70
T + V	72.14 ± 0.58	71.76 ± 0.49	71.90 ± 0.51
A + V	66.90 ± 0.62	66.57 ± 0.53	66.75 ± 0.52
T + A	60.70 ± 0.45	60.43 ± 0.63	60.57 ± 0.52
T	69.59 ± 0.54	69.59 ± 0.69	69.54 ± 0.55
V	61.40 ± 0.53	61.16 ± 0.82	61.34 ± 0.59
A	56.64 ± 0.49	56.08 ± 0.66	56.44 ± 0.52

4.2.3 Parameter Analysis and Complexity

In order to provide more convincing evidence to prove the superiority of the proposed model, we have shown and compared the number, complexity and running time of parameters of our model and the other six most advanced baselines in Table 6, namely UPB-MTL, RCNN RoBERTa, QFNN, A-MTL, M2Seq2Seq, SRLM. In addition, the QUIET model proposed by Liu et al. [28] can achieve high performance by introducing complex frameworks such as quantum probability theory, but its parameters are relatively large.

Table 6: Model eficiency comparison.

Model	**Parameters (Millions)**	**Training Time (h)**
UPB-MTL	4.1 M	1.7
RCNN-RoBERTa	125 M	2.1
QFNN	1.7 M	0.5
A-MTL	2.0 M	0.4
M2Seq2Seq	3.3 M	1.2
SRLM	3.2 M	1.2
Ours	1.97 M	0.4

As can be seen from the data in Table 6, compared with the RCNN-RoBERTa model, most models have a smaller number of parameters. Specifically, the number of parameters in our model is 1.97 M, second only to the 1.7 M of the lightest QFNN among all the compared models, and is lower than that of A-MTL. However, our model demonstrates excellent performance. This proves that the superiority of our model does not stem from simple parameter stacking, but rather from its internal rational architecture design, which has successfully achieved a breakthrough in high-performance under the constraint of low complexity.

4.2.4 Ablation Study

In order to further analyze the role of each module in model improvement, we conducted ablation experiments with Mustard dataset as an example: (1) All: reserve all core modules of the model; (2) No Quantum interference fusion: remove the quantum interference fusion module, and use the cross attention mechanism to achieve cross modal fusion; (3) No complex value embedding: remove the complex value embedding module and directly use the original sequence features for subsequent calculation; (4) No Module Asymmetry: remove the trainable parameters in the quantum interference module, and maintain the symmetry of the original quantum interference. (5) Real-valued version with matched params: Build a real-valued model variant with matched parameter scale, remove all complex-valued operations, and conduct all modal processing and fusion with pure real-valued features. (6): No Phase: Retain only modal amplitude features, fix the imaginary part to zero, and conduct all subsequent processing solely with amplitude information.

As shown in the ablation experiment results in Table 7, the quantum interference fusion module makes the greatest contribution to the overall performance, because it effectively captures the "non classical association" among text, visual, and audio multimodes, providing a key multimodal interaction feature for sarcasm detection. Then there are complex value embedding and asymmetric interaction between modes. Complex value embedding captures the uncertainty and long-term correlation of semantic units in modes through the "amplitude phase" two-dimensional modeling; The asymmetric interaction between modes adapts to the asymmetric dependency of "semantic constraint ambiguity supplement" between multimodes, and accurately describes the differential interaction logic between modes.

Table 7: Ablation study.

Model	F1	F1 (weighted)	ACC
All	78.31 ± 0.83	78.31 ± 0.86	78.46 ± 0.70
No Quantum interference fusion	72.36 ± 0.81	72.04 ± 0.90	72.16 ± 0.84
No complex value embedding	74.51 ± 0.50	74.47 ± 0.52	74.48 ± 0.52
No Module Asymmetry	76.69 ± 0.48	76.49 ± 0.56	76.54 ± 0.52
Real-valued version with matched params	74.59 ± 0.79	74.05 ± 0.95	74.23 ± 0.84
No Phase	75.08 ± 0.57	74.95 ± 0.49	75.0 ± 0.82

4.2.5 Validating Asymmetric Multimodal Fusion Design

In order to verify the effectiveness of asymmetric multimodal fusion design, we conducted the following experiments: (1) All (2) Remove V-A (3) Remove A-V (4) Remove V-T (5) Remove T-V (6) Remove T-A (7) Remove A-T.

As shown in the ablation experiment results in Table 8. Text plays a leading role in this multimodal fusion framework, and its two-way interaction with vision is the core support to maintain the model performance. When removing text and vision related interaction paths, the impact on model performance is most significant, which is far more than the consequences of removing other paths; In contrast, removing audio related paths will only cause small fluctuations in performance. This result fully confirms the leading role of text and the importance of vision as a key collaborative mode. The two-way interaction between the two is the key to ensure the effectiveness of the model in the task.

Table 8: Validating asymmetric multimodal fusion design.

Mustard			
Model	**F1**	**F1 (weighted)**	**ACC**
All	78.31 ± 0.83	78.31 ± 0.86	78.46 ± 0.70
Remove V-A	76.05 ± 0.53	76.03 ± 0.51	76.03 ± 0.52
Remove A-V	76.36 ± 0.84	76.26 ± 0.84	76.29 ± 0.84
Remove V-T	75.84 ± 0.64	75.75 ± 0.58	75.77 ± 0.59
Remove T-V	74.88 ± 0.63	74.70 ± 0.58	74.74 ± 0.60
Remove T-A	75.63 ± 0.55	75.48 ± 0.51	75.52 ± 0.52
Remove A-T	77.32 ± 0.83	77.31 ± 0.83	77.31 ± 0.84

4.3 Error Analysis

Table 9 clearly reveals the asymmetry of the text image auxiliary relationship in multimodal sarcasm detection, which is also the core reason for the misjudgment of the traditional symmetric fusion model in satire detection. In cases (1) and (3), the text itself carries strong satirical signals, and the pictures are only used to supplement the context or reinforce the tone; If the symmetrical model gives too much weight to the image, the ironic core of the text may be weakened due to scene deviation, leading to misjudgment. In contrast, in (2) and (4) cases, the text is neutral and has no satirical features, and the satirical logic is completely provided by the contrast between the scene and emotion of the picture; The symmetrical model, because of the equal weight of text and pictures, will be misjudged as non satirical because the text has no satirical signal, ignoring the dominant role of pictures.

Table 9: Asymmetric cases.

 (1) I'm just inferring this is a couch because the evidence suggests the coffee table is having a tiny garage sale.	 (2) If you don't like this stuff, let's go next door and build her a bear.
 (3) Well, I'm sorry, too, but there's just no room for you in my wallet.	 (4) Oh I am sorry, do you need a break?

5 Conclusions and Future Work

With the increasing richness and diversification of multimedia data, multimodal sarcasm identification will play an increasingly important role in the development of multimedia. In this paper, a complex value fusion framework based on quantum heuristic is proposed to solve the problems of insufficient semantic expression within modes and inaccurate modeling of interaction between modes in multimodal sarcasm detection tasks. By introducing the fusion mechanism of feature representation and asymmetric quantum

interference in complex Hilbert space, the framework effectively enhances the expression ability of satirical features and the modeling ability of cross modal interaction, and proves its effectiveness on Mustard and Memotion datasets.

Future research will be devoted to deepening and expanding this framework from multiple dimensions: richer quantum cognitive mechanisms such as quantum entanglement and quantum measurement can be introduced to model deeper nonclassical correlations between modes and dynamically make the decision-making process; The model proposed in this paper can be applied to humor detection, false information recognition and other related tasks to explore its potential to solve multimodal inconsistency problems.

Acknowledgement: Not applicable.

Funding Statement: This research was supported by Youth Innovation Team Project, Scientific Research Program of Shaanxi Provincial Department of Education (No. 25JP070), the Natural Science Foundation of China (Nos. 62072362, 12101479), Natural Science Basis Research Plan in Shaanxi Province of China (Nos. 2021JQ-660 and 2024JC-YBMS-531), Shaanxi Provincial Innovation Capacity Support Programme Project (No. 2024ZC-KJXX-034), and Xi'an Major Scientific and Technological Achievements Transformation Industrialization Project (No. 23CGZHCYH0008).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Dong Zhang and Lianhe Shao; methodology, Dong Zhang and Lianhe Shao; software, Dong Zhang; validation, Dong Zhang and Quanli Gao; formal analysis, Dong Zhang and Xihan Wang; investigation, Dong Zhang and Lianhe Shao; resources, Dong Zhang, Weijie Xu and Quanli Gao; data curation, Dong Zhang; writing—original draft preparation, Dong Zhang; writing—review and editing, Dong Zhang, Lianhe Shao, Weijie Xu, Xihan Wang and Quanli Gao; supervision, Lianhe Shao; project administration, Dong Zhang, Lianhe Shao, Weijie Xu and Quanli Gao; funding acquisition, Lianhe Shao, Weijie Xu, Quanli Gao and Xihan Wang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report.

References

1. Zhang Y, Wang J, Liu Y, Rong L, Zheng Q, Song D, et al. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations. Inf Fusion. 2023;93(5):282–301. doi:10.1016/j.inffus.2023.01.005.
2. Tiwari P, Zhang L, Qu Z, Muhammad G. Quantum Fuzzy Neural Network for multimodal sentiment and sarcasm detection. Inf Fusion. 2024;103(2):102085. doi:10.1016/j.inffus.2023.102085.
3. Zhang Y, Yu Y, Wang M, Huang M, Hossain MS. Self-adaptive representation learning model for multi-modal sentiment and sarcasm joint analysis. ACM Trans Multimed Comput Commun Appl. 2024;20(5):1–17. doi:10.1145/3635311.
4. Veale T, Hao Y. Detecting ironic intent in creative comparisons. In: ECAI 2010. Amsterdam, The Netherlands: IOS Press; 2010. p. 765–70. doi:10.3233/978-1-60750-606-5-765.
5. Maynard D, Greenwood MA. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In: 9th International Conference on Language Resources and Evaluation; 2014 May 26–31; Reykjavik, Iceland.
6. Schifanella R, de Juan P, Tetreault J, Cao L. Detecting sarcasm in multimodal social platforms. In: Proceedings of the 24th ACM International Conference on Multimedia; 2016 Oct 15–19; Amsterdam, The Netherlands. p. 1136–45. doi:10.1145/2964284.2964321.

7. Srivastava H, Varshney V, Kumari S, Srivastava S. A novel hierarchical BERT architecture for sarcasm detection. In: Proceedings of the Second Workshop on Figurative Language Processing; 2020 Jul 9; Online. Stroudsburg, PA, USA: ACL; 2020. p. 93–7. doi:10.18653/v1/2020.figlang-1.14.
8. Wang X, Sun X, Yang T, Wang H. Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data. In: Proceedings of the First International Workshop on Natural Language Processing Beyond Text; 2020 Nov 20; Online. p. 19–29. doi:10.18653/v1/2020.nlpbt-1.3.
9. Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 2506–15. doi:10.18653/v1/P19-1239.
10. Wu Y, Lin Z, Zhao Y, Qin B, Zhu LN. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021 Aug 1–6; Online. p. 4730–8. doi:10.18653/v1/2021.findings-acl.417.
11. Ding N, Tian SW, 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.
12. Pan H, Lin Z, Fu P, Qi Y, Wang W. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In: Findings of the Association for Computational Linguistics: EMNLP 2020; 2020 Nov 16–20; Stroudsburg, PA, USA: ACL; 2020. p. 1383–92. doi:10.18653/v1/2020.findings-emnlp.124.
13. 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. p. 1767–77. doi:10.18653/v1/2022.acl-long.124.
14. Qiao Y, Jing L, Song X, Chen X, Zhu L, Nie L. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. Proc AAAI Conf Artif Intell. 2023;37(8):9507–15. doi:10.1609/aaai.v37i8.26138.
15. Wen C, Jia G, Yang JDIP. Dual incongruity perceiving network for sarcasm detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 2540–50. doi:10.1109/CVPR52729.2023.00250.
16. Bedi M, Kumar S, Akhtar MS, Chakraborty T. Multi-modal sarcasm detection and humor classification in code-mixed conversations. IEEE Trans Affect Comput. 2023;14(2):1363–75. doi:10.1109/taffc.2021.3083522.
17. Dubey P, Dubey P, Bokoro PN. Unpacking sarcasm: a contextual and transformer-based approach for improved detection. Computers. 2025;14(3):95. doi:10.3390/computers14030095.
18. Lee H, Yu Y, Kim G. Augmenting data for sarcasm detection with unlabeled conversation context. In: Proceedings of the Second Workshop on Figurative Language Processing; 2020 Jul 9; Online. Stroudsburg, PA, USA: ACL; 2020. p. 12–7. doi:10.18653/v1/2020.figlang-1.2.
19. Gao X, Bansal S, Gowda K, Li Z, Nayak S, Kumar N, et al. AMuSeD: an attentive deep neural network for multimodal sarcasm detection incorporating bimodal data augmentation. IEEE Trans Affect Comput. 2026;17(1):900–12. doi:10.1109/taffc.2025.3639406.
20. Kak SC. Quantum neural computing. In: Advances in imaging and electron physics. Amsterdam, The Netherlands: Elsevier; 1995. p. 259–313. doi:10.1016/s1076-5670(08)70147-2.
21. 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.
22. Liu Y, Zhang Y, Li Q, Wang B, Song D. What does your smile mean? Jointly detecting multi-modal sarcasm and sentiment using quantum probability. In: Findings of the Association for Computational Linguistics: EMNLP 2021; 2021 Nov 7–11; Punta Cana, Dominican Republic. Stroudsburg, PA, USA: ACL; 2021. p. 871–80. doi:10.18653/v1/2021.findings-emnlp.74.
23. Yan K, Lai P, Zheng X, Yang Y, Ren Y, Badarch T, et al. Quantum-inspired neural networks with stochastic dynamics for multimodal sentiment analysis and sarcasm detection. Eng Appl Artif Intell. 2026;163(24):112923. doi:10.1016/j.engappai.2025.112923.
24. Yan K, Lai P, Yang Y, Ren Y, Badarch T, Chen Y, et al. Quantum-inspired multimodal fusion with Lindblad master equation for sentiment analysis. Neurocomputing. 2025;648(24):130710. doi:10.1016/j.neucom.2025.130710.

25. Vlad GA, Zaharia GE, Cercel DC, Chiru CG, Trausan-Matu S. UPB at SemEval-2020 task 8: joint textual and visual modeling in a multi-task learning architecture for memotion analysis. arXiv:2009.02779. 2020.
26. Potamias RA, Siolas G, Stafylopatis AG. A transformer-based approach to irony and sarcasm detection. Neural Comput Appl. 2020;32(23):17309–20. doi:10.1007/s00521-020-05102-3.
27. Chauhan DS, SR D, Ekbal A, Bhattacharyya P. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. Stroudsburg, PA, USA: ACL; 2020. p. 4351–60. doi:10.18653/v1/2020.acl-main.401.
28. Liu Y, Zhang Y, Song D. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. IEEE Trans Affect Comput. 2024;15(1):326–41. doi:10.1109/TAFFC.2023.3279145.