What Does Your Smile Mean? Jointly Detecting Multi-Modal Sarcasm and Sentiment Using Quantum Probability

Yaochen Liu\textsuperscript{1}, Yazhou Zhang\textsuperscript{2,3,}\ast, Qiuchi Li\textsuperscript{4}, Benyou Wang\textsuperscript{5}, Dawei Song\textsuperscript{1,†}

\textsuperscript{1}Beijing Institute of Technology
\textsuperscript{2}Zhengzhou University of Light Industry
\textsuperscript{3}State Key Lab. for Novel Software Technology, Nanjing University
\textsuperscript{4}University of Copenhagen, \textsuperscript{5}University of Padua
\{yaochen, dwsong\}@bit.edu.cn, yzzhang@zzuli.edu.cn, qiuchi.li@di.ku.dk, wang@dei.unipd.it

Abstract

Sarcasm and sentiment embody intrinsic uncertainty of human cognition, making joint detection of multi-modal sarcasm and sentiment a challenging task. In view of the advantages of quantum probability (QP) in modeling such uncertainty, this paper explores the potential of QP as a mathematical framework and proposes a QP driven multi-task (QPM) learning framework. The QPM framework involves a complex-valued multi-modal representation encoder, a quantum-like fusion network and a quantum measurement mechanism. Each multi-modal (e.g., textual, visual) utterance is first encoded as a quantum superposition of a set of basis terms using a complex-valued representation. Then, the quantum-like fusion network leverages quantum state composition and quantum interference to model the contextual interaction between adjacent utterances and the correlations across modalities respectively. Finally, quantum incompatible measurements are performed on the multi-modal representation of each utterance to yield the probabilistic outcomes of sarcasm and sentiment recognition. Experimental results show the state-of-the-art performance of our model.

1 Introduction

Multi-modal sarcasm and sentiment analysis, as a challenging problem, has attracted an increasing attention in the recent literature (Cai et al., 2019; Pan et al., 2020). Sarcasm is a subtle form of human language that intends to express criticism, humor or mock sentiments by means of hyperbole, figuration, etc (Castro et al., 2019). The literal meaning of an ironic expression differs from its real implication, which can completely flip the polarity of sentiment. Hence, sentiment comes into view and tightly couples with sarcasm in that one helps the understanding of the other. Consequently, jointly detecting sarcasm and sentiment would bring benefits to each other.

Judging sarcasm and sentiment of human language, e.g., an utterance in a conversation, involves intrinsically uncertain human cognition processes (Carroll and Carroll, 1999). The uncertainty is rooted on the spontaneity of human subjective activities, where the generation of sarcasm and sentiment is often spontaneous and intuitive without a rational reasoning process. Meanwhile, human language is multi-modal in nature, involving multi-modal (e.g., textual and visual) features that interact with each other and introduce extra cognitive complexity. Thus, it is essential to study sarcasm and sentiment from a general cognitive perspective.

Motivated by recent success in using quantum probability (QP) as a formal framework for modeling the intrinsic uncertainty in human cognition, we take the first step towards using QP to solve the joint multi-modal sarcasm and sentiment analysis problem. Originally as the mathematical foundation of quantum mechanics that describes the behaviors of particles, QP has been employed to formalize the uncertainty in various macro-tasks such as semantic analysis (Bruza et al., 2009; Uprety et al., 2020), question answering (Zhang et al., 2018a; Li et al., 2019) and sentiment classification (Zhang et al., 2020; Gkoumas et al., 2021), with verified effectiveness and advantages. Different from these existing approaches, at the heart of our work are quantum inspired modeling of multi-modal fusion in conversational context and exploring the inter-task correlations via quantum incompatible measurement.

The reasons to use QP are four fold: (1) QP is advantageous in modeling the uncertainty in human cognition because it introduces the concept of complex probability amplitude, and models an utterance as a quantum superposition of basis words or pixels; (2) Quantum interference embodies a...
non-linear fusion of multi-modal features, due to an interference term for modeling two decision paths (e.g., textual and visual modalities) interfering with each other in reaching a final decision (e.g., sarcasm judgment); (3) Quantum contextuality reflects the intra-modality contextual interaction as quantum composition; (4) Quantum incompatible measurement describes the correlations across multiple tasks. Since sarcasm and sentiment are tightly coupled, we thus argue that they are incompatible, i.e., judging one will affect the judgment of the other. To sum up, we can intuitively discover some commonality between QP and multi-modal sarcasm and sentiment analysis, and benefit from the unified and principled mathematics of QP. A detailed formal explanation is provided in Sec. 3.

In this paper, we propose a QP driven multi-task (QPM) learning framework. Specially, QPM involves a complex-valued multi-modal representation encoder, a quantum-like fusion network and a quantum measurement mechanism. First, inspired by (Li et al., 2019), each modality of utterance is described as a quantum superposition of a set of basis semantic units and represented by a complex-valued embedding. Then, we propose a quantum-like fusion network that leverages quantum state composition and quantum interference to capture intra-modal contextuality and inter-modal incongruity. The contextuality is described as the contextual interaction between adjacent utterances, which is mathematically encapsulated in a density matrix. The inter-modal incongruity is handled at the feature level with a quantum interference-like fusion approach. Finally, since all the information contained in one system is represented by the probability distribution of quantum measurement results, the final multi-modal features can be extracted via quantum incompatible measurement, while these features are passed to a fully connected layer to yield sarcasm and sentiment predictions.

Extensive empirical results on two benchmark datasets, MUS cran and Memotion, show that the effectiveness of QPM over state-of-the-art baselines. The major innovations of the work are:

- The first QP driven multi-task learning framework for joint multi-modal sarcasm and sentiment analysis.
- A quantum-like fusion network for modeling intra-modality contextuality and inter-modality incongruity.
- A quantum incompatible measurement approach capturing inter-task dependency.

2 Quantum Probability Preliminaries

Quantum Superposition and Density Matrix. The mathematical base of quantum probability is established on a complex Hilbert Space, denoted as $\mathcal{H}$. A quantum state vector $|u\rangle$ is expressed as a ket $|u\rangle$, its transpose is expressed as a bra $\langle u|$. The inner product and outer product of two state vectors $|u\rangle$ and $|v\rangle$ are denoted as $\langle u|v\rangle$ and $|u\rangle\langle v|$. Quantum superposition states that a pure quantum state can be in multiple mutually exclusive basis states simultaneously, with a probability distribution until it is measured. A quantum mixture of states gives rise to a mixed state represented by a density matrix, $\rho = \sum_i p_i |u_i\rangle\langle u_i|$, where $p_i$ denotes the probability distribution of each pure state.

Quantum Interference. In the double-slit experiment, two paths interfering with each other affects the probability distribution of the particle reaching the final position of the detection screen. We use the wave function $\varphi(x)$ to interpret this behavior. The wave function represents the probability amplitude of a particle be at a position $x$, and the square of the wave function represents the possibility. The state of the photon is in a quantum superposition of the state of path 1 and path 2, which is formulated as: $\varphi_p(x) = \alpha \varphi_1(x) + \beta \varphi_2(x)$, where $\varphi_1(x)$ and $\varphi_2(x)$ are the wave function of path 1 and path 2. $\alpha$ and $\beta$ are complex numbers. Its probability is:

$$P(x) = |\varphi_p(x)|^2 = |\alpha \varphi_1(x) + \beta \varphi_2(x)|^2$$

$$= |\alpha \varphi_1(x)|^2 + |\beta \varphi_2(x)|^2 + 2|\alpha \beta \varphi_1(x) \varphi_2(x)| \cos \phi$$

(1)

where $\phi$ is the interference angle. $I = 2|\alpha \varphi_1(x) \beta \varphi_2(x)| \cos \phi$ is the interference term, which describes the interaction between two paths.

Quantum Measurement. Quantum measurement is described by a set of measurement operators acting on the state space of the system being measured $\{M_m\}$, where $m$ represents the possible measurement outcomes. Suppose the quantum system is in a state of $|u\rangle$, then the probability to obtain the outcome $m$ after the measurement is $p(m) = \langle u|M_m^\dagger M_m|u\rangle$. The Gleason’s Theorem (Sordoni et al., 2013) has proven the existence of a mapping function $M(|u\rangle\langle u|) = tr(\rho|u\rangle\langle u|)$ for any event $|u\rangle\langle u|$. 

872
3 Theoretical Justification of the Proposed QPM Framework

Based on the general QP and a few previous studies (Wang et al., 2019; Li et al., 2019), this section proposes theoretical justification of our QPM framework in the form of four claims.

Claim 1 Quantum probability is more general to capture the uncertainty in human language.

Assume \( z(x) \) represents a complex probability amplitude of an event \( x \), where \( z(x) = re^{i\theta} \). QP defines the modulus square of this complex probability amplitude to represent a classical probability \( p(x) = |z(x)|^2 = r^2 \). It defines a many-to-one relationship between complex probability amplitude and probability.

For example, the probability of a word \( w \) is 0.5, i.e., \( p(x = w) = \frac{1}{2} \), then the corresponding probability amplitude may be \( z(x = w) = \frac{\sqrt{2}}{2} e^{i\frac{\pi}{4}} \) or \( z(x = w) = \frac{\sqrt{2}}{2} e^{i\frac{3\pi}{4}} \), etc. The amplitude \( r \) links to the probability, while the phase \( \theta \) may be associated with hidden sentiment or sarcasm orientations. The reasons are: (1) by using this formulation, two antonym words could have similar amplitudes but they may have different sentimental polarities represented in the phase term. (2) Words often carry multiple dimensions (e.g., semantic and sentiment) of information. It is reasonable to use amplitude-phase format to model the semantic and sentiment jointly. Then, an utterance could be represented in an amplitude-phase manner.

Claim 2 Quantum interference embodies a non-linear multi-modal fusion.

Quantum interference describes a phenomenon that two propagation paths (e.g., textual and visual channels) interfering with each other affects the probability distribution of a particle (e.g., the author’s attitude). Assume \( z(x) \) represents a complex probability amplitude of the modality \( x \), the probability amplitude of multi-modality that consists of two modalities \( x_1, x_2 \) can be formalized as:

\[
z_3(x_3) = \alpha z_1(x_1) + \beta z_2(x_2)
\]

where \( \alpha \) and \( \beta \) are complex coefficients. The probabilities of \( x_1 \) and \( x_2 \) are measured as:

\[
p(x_1) = |\alpha|^2 |z_1(x_1)|^2, \quad p(x_2) = |\beta|^2 |z_2(x_2)|^2
\]

We can derive the probability of multi-modality:

\[
p(x_3) = |z_3(x_3)|^2 = |\alpha z_1(x_1) + \beta z_2(x_2)|^2
= p(x_1) + p(x_2) + 2\sqrt{p(x_1)p(x_2)}\cos\phi
= p(x_1) + p(x_2) + \sqrt{p(x_1)p(x_2)}(e^{i\phi} + e^{-i\phi})
\]

(4)

Hence, the probability of multi-modality is a non-linear combination of the probabilities of two uni-modalities, with an interference term determined by the relative phase \( \phi \). This provides a higher level of abstraction (Jiang et al., 2020; Li et al., 2021).

Claim 3 Quantum composition captures the contextuality between utterances.

Quantum contextuality describes the results of measurements on a particle depending on the measurement environment. This intuitively reflects the phenomena that the sarcastic and sentimental states of an utterance are decided by its contexts. Assume \( u_i \) and \( u_j \) represent two adjacent utterances in a conversation, each of which is made up of two basis words:

\[
\begin{align*}
|u_i⟩ &= \alpha_1|w_1⟩ + \beta_1|w_2⟩, \quad |u_j⟩ = \alpha_2|w_1⟩ + \beta_2|w_2⟩ \\
\end{align*}
\]

(5)

The contextual interaction between utterances \( u_i \) and \( u_j \) constructs the state space of a composite system \( H_{u_i,u_j} \), which is defined as a tensor product of the individual state spaces \( |u_i⟩ \) and \( |u_j⟩ \):

\[
H_{u_i,u_j} = |u_i⟩ \otimes |u_j⟩ = \alpha_1\alpha_2|w_1w_1⟩ + \alpha_1\beta_2|w_1w_2⟩ + \beta_1\alpha_2|w_2w_1⟩ + \beta_1\beta_2|w_2w_2⟩
\]

(6)

Eq. 6 shows that the composition system consisting of utterances embodies the correlations between words, which inspires us to model the contextuality by a “global to local” way (Zhang et al., 2018b).

Claim 4 Quantum incompatible measurement describes the correlations across multi-tasks.

Given two sets of \( G \) measurement operators for sarcasm and sentiment observables, \( M_{sar} = \{M_{\gamma}^{sar}\}_{\gamma=1}^G \), \( M_{sen} = \{M_{\delta}^{sen}\}_{\delta=1}^G \). If any cross-task pair of measurement operators satisfy the commutation rule\(^1\), i.e., \( [M_{\gamma}^{sar},M_{\delta}^{sen}] = 0 \) for all \( \gamma \) and \( \delta \), then the sarcasm and sentiment observables are called compatible, otherwise we say they are incompatible (Designolle et al., 2019). Here, sarcasm and sentiment are tightly intertwined and the judgment on one may affect the other. Thus we intuitively argue that they are incompatible, and check whether our hypothesis is tenable in the experiments (c.f. Sec. 5.8). We introduce quantum

\(^1\)\( [M_{\gamma}^{sar},M_{\delta}^{sen}] = M_{\gamma}^{sar}M_{\delta}^{sen} - M_{\delta}^{sen}M_{\gamma}^{sar} = 0 \)
relative entropy to quantitatively analyze the inter-task correlation, and help measure specific degree of correlation across different tasks.

4 Methodology

4.1 Task Definition and Overall Network

Task Definition. Suppose the dataset has $L$ multimodal samples. The $i^{th}$ sample $X_i$ is represented as $\{X^i = (C_i, U^i, Y^i), \}$, where $C^i$, $U^i$, $Y^i$ denote the $i^{th}$ conversational context, the multi-modal utterance and the label respectively, and $i \in [1, 2, ..., L]$. Both the context and the multi-modal utterance consist of textual and visual modalities, i.e., $C_i = (C_t^i, C_v^i)$, $U^i = (U_t^i, U_v^i)$.

Now, the task of multi-modal sarcasm and sentiment detection can be formulated as:

$$\zeta = \prod_{i} p \left( Y^i | C^i, U^i, \Theta \right)$$

where $\Theta$ represents the parameter set.

Overall Network. The overall architecture of the QPM framework is shown in Figure 1. (1) The $i^{th}$ textual utterance and its visual counterpart are represented by complex-valued embeddings, denoted as $|u_t^i\rangle$ and $|u_v^i\rangle$. (2) Then, $|u_t^i\rangle$ and $|u_v^i\rangle$ are fed into the quantum composition layer to capture the contextuality, where the results are encapsulated in two density matrices $\rho_{text}$ and $\rho_{img}$. (3) We then fuse $\rho_{text}$ and $\rho_{img}$ for obtaining a multi-modal representation via the quantum interference layer. (4) We extract the final sarcastic and sentimental features via quantum incompatible measurement, and feed these features into a fully connected softmax layer to yield sarcasm and sentiment predictions.

4.2 Complex-valued Textual and Visual Embedding

Inspired by Li and Wang’s work (Li et al., 2019), for textual modality, an utterance can be seen as a collection of words. We assume that the textual Hilbert space $\mathcal{H}_t$ is spanned by a set of orthogonal basis states $\{|w^i_j\rangle\}_{j=1}^n$. With words as the basic semantic unit, the $j^{th}$ word $w_j$ can be used as the basis state $|w^i_j\rangle$, represented by one-hot encoding, i.e., the $j$-th element being 1 and 0s elsewhere.

Then, we regard the $i^{th}$ target utterance $u_t^i$ as a quantum superposition of a set of basis words $\{|w^i_1\rangle, |w^i_2\rangle, ..., |w^i_n\rangle\}$, which is formulated as:

$$|u_t^i\rangle = \sum_{j=1}^{n} z_j^i |w^i_j\rangle, \quad z_j^i = r^i_j e^{i\theta_j^i}$$

where $n$ is the number of words in the utterance. $z_j^i$ is a complex probability amplitude expressed in the polar form. $i$ is the imaginary number. $r_j^i$ is the modulus of the complex number, termed amplitude. $\theta_j^i \in [-\pi, \pi]$ is the argument (phase) of $z_j^i$.

We construct the complex-valued vector of the $\zeta_t^{th}$ utterance, by associating the amplitude $r$ with the semantic knowledge and the phase $\theta$ with the pre-assigned sentiment orientation, i.e., $|u_t^i\rangle = \left( r_1 \cos \theta_1, r_1 \sin \theta_1, ..., r_n \cos \theta_n \right)^T$.

For visual modality, the low-level visual features are seen as the basic unit. We assume that the visual Hilbert space $\mathcal{H}_v$ is spanned by a set of orthogonal basis visual features $\{|w^i_j\rangle\}_{j=1}^k$, where the visual part of the target utterance is represented as $|u_v^i\rangle$.

The textual and visual embeddings of $i^{th}$ contextual utterance, $|c_t^i\rangle$ and $|c_v^i\rangle$, can be calculated in the same way.

4.3 Learning Intra-modality Contextuality with the Quantum Composition Layer

Treating the target multimodal utterance as a quantum system, its contexts as the surrounding environments, we propose a quantum composition layer to learn the intra-modality contextuality.

For text, given that the target utterance $|u_t^i\rangle$ and its contexts $\{|c^i_1\rangle, ..., |c^i_k\rangle\}$, the contextual interaction between them constructs a textual composite system $\Psi_t^{i,k}$, which is given by the tensor product of individual utterance embeddings. We aim to learn both long and short range contextual interactions, by constructing multiple composite systems with a variable number of contexts. The $\lambda^{th}$ composite system is computed as:

$$|\Psi_t^{i,\lambda}\rangle = |u_t^i\rangle \otimes |c^i_1\rangle \otimes |c^i_2\rangle \otimes ..., \otimes |c^i_\lambda\rangle$$

where $\lambda \in [1, k]$. We can build $k$ composite systems for $k$ context utterances, i.e., $\Psi_{t,k} = \{\Psi_t^{1,k}, \Psi_t^{2,k}, ..., \Psi_t^{k,k}\}$.

These $k$ composite systems are mathematically encapsulated in a textual density matrix $\rho_{text}$, to obtain the representation of the target utterance $u_t^i$.

$$\rho_{text} = \sum_{\lambda=1}^{k} p_\lambda |\Psi_t^{i,\lambda}\rangle \langle \Psi_t^{i,\lambda}|$$

where $p_\lambda$ represents the weights to be learned during training. The density matrix unifies the target utterance and its contexts.

For the visual part, we also build $k$ composition system for $k$ visual contexts, i.e., $\Psi_v^{i,k} =$
\[
\left\{ |\Psi_1^x\rangle, |\Psi_2^x\rangle, \ldots, |\Psi_k^x\rangle \right\},
\]
and obtain the visual density matrix \(\rho_{\text{img}}\) using Eq. 10.

Then, textual and visual density matrices \(\rho_{\text{text}}\) and \(\rho_{\text{img}}\) are flattened into two vectors \(|f_1\rangle\) and \(|f_2\rangle\) for multi-modal fusion via quantum interference.

### 4.4 Quantum Interference-like Fusion Layer

Based on Eq. 2, 3 and 4, we argue that the subjective attitude of a speaker is in a quantum superposition-like of textual and visual representations, expressed as:

\[
\text{z}_p(x) = \alpha \text{z}_t(x) + \beta \text{z}_v(x)
\]  

(11)

where \(\text{z}_t(x)\) and \(\text{z}_v(x)\) represent the complex probability amplitudes of textual and visual representations. \(f_t(x) = |\alpha|^2|\text{z}_t(x)|^2\) and \(f_v(x) = |\beta|^2|\text{z}_v(x)|^2\) represent the corresponding probability distributions. The probability distribution of multi-modal representation is then written as:

\[
f_p(x) = f_t(x) + f_v(x) + 2\sqrt{f_t(x)f_v(x)} \cos \phi
\]  

(12)

where \(x\) is the \(x^{th}\) feature component of the multi-modal representation \(|f_p\rangle\). \(I = 2\sqrt{f_t(x)f_v(x)} \cos \phi\) is the interference item. \(|f_p\rangle = (f_p(x_1), f_p(x_2), \ldots, f_p(x_n))^T\) represent the multi-modal fused features.

### 4.5 Quantum Measurement Layer

In QP, the properties of a system (e.g., an utterance’s sarcastic information) can be depicted by the probability distribution of the measurement outcomes. The multi-modal representation \(|f_p\rangle\) is shared across the two branches of our proposed QPM, and we propose to perform a sequence of quantum incompatible measurements on \(|f_p\rangle\) for obtaining the sarcastic and sentimental probabilistic features \(\tilde{m}^{\text{sar}}\) and \(\tilde{m}^{\text{sen}}\).

Specifically, two sets of measurement operators \(M^{\text{sar}} = \{M^{\text{sar}}_\gamma\}_\gamma, M^{\text{sen}} = \{M^{\text{sen}}_\delta\}_\delta\) are pre-defined, each constructed by the outer product of the corresponding measurement vector \(|E_\gamma\rangle\) or \(|A_\delta\rangle\), i.e., \(M^{\text{sar}}_\gamma = |E_\gamma\rangle\langle E_\gamma|, M^{\text{sen}}_\delta = |A_\delta\rangle\langle A_\delta|\). The probability distribution over the measurement outcomes can be computed as:

\[
\tilde{m}^s = tr \left( (M^s)^\dagger M^s |f_p\rangle \langle f_p| \right), \quad s \in \{\text{sar}, \text{sen}\}.
\]

### 4.6 Dense Layer

The sarcastic and sentimental outcomes \(\tilde{m}^{\text{sar}}, \tilde{m}^{\text{sen}}\) are forwarded through a fully connected layer and the softmax function to yield the sarcasm and sentiment predictions. We use cross entropy with L2 regularization as the loss functions \(\zeta^{\text{sar}}\) and \(\zeta^{\text{sen}}\), and jointly minimize them with different weights, e.g., \(\zeta = w^{\text{sar}}\zeta^{\text{sar}} + w^{\text{sen}}\zeta^{\text{sen}}\). We receive gradients of error from two branches and accordingly adjust the weights.

## 5 Experiments and Analysis

### 5.1 Experiment Settings

**Datasets.** We choose benchmark datasets that have textual and visual modalities with both sarcasm and sentiment labels. Only the extended version of MUSTARD (MUSTARD_ext for short)\(^2\) (Chauhan 2019).

\(^2\)http://www.iitp.ac.in/ai-nlp-ml/resources.html
A wide range of state-of-the-art baselines are included for comparison. They are:

**SVM+BERT** (Devlin et al., 2019): It represents the textual utterances using BERT vectors and standard hyperparameter settings. We also concatenate the contextual features.

**RCNN-RoBERTa** (Potamias et al., 2020): It utilizes pre-trained RoBERTa vectors combined with a RCNN in order to capture contextual information.

**EfficientNet** (Tan and Le, 2019): It uses a compound scaling method to create different models, which has achieved state-of-the-art performance on the ImageNet challenge.

**UPB-MTL** (Vlad et al., 2020): It is a multimodal multi-task learning architecture that combines ALBERT for text encoding with VGG-16 for image representation.

**QMSA** (Zhang et al., 2018c): It first extracts visual and textual features using density matrices, and feeds them into the SVM classifier.

**A-MTL framework** (Chauhan et al., 2020): It proposes an attention based multi-task model to simultaneously analyse sentiment, emotion and detect sarcasm.

### 5.3 Comparative Analysis

The experimental results are summarized in Table 3. Text-QPM and Image-QPM, which are single-modality variants of QPM, do not perform well, demonstrating that text or visual modalities cannot be treated independently for multi-modal sarcasm and sentiment detection. The proposed QPM model achieves the best micro-F1 of 77.53% as compared to 76.57% of the state-of-the-art system (i.e., A-MTL) on MUSTARD$_{ext}$. QPM achieves a micro-F1 of 61.39% as compared to 59.85% of A-MTL on Memotion. The results show that the proposed QPM framework leverages the advantages of QP in modeling the uncertainty in human language. We attribute the main improvements to the proposed QPM framework leveraging the advantages of QP in modeling the uncertainty in human language. We attribute the main improvements to the proposed QPM framework leveraging the advantages of QP in modeling the uncertainty in human language.
We perform an ablation study to further study the effectiveness of different components of QPM: (1) QPM-Trad that does not consider the complex embedding, i.e., replacing utterance embeddings with their real counterparts; (2) QPM-Speaker Independent without modeling contextuality; (3) QPM-Concat that replaces the quantum interference-like fusion layer with multi-modal concatenation; (4) QPM-Trad that replaces quantum incompatible measurements with traditional softmax layers.

The results in Table 6 show that quantum incompatible measurement contributes the most to overall performance, as it effectively captures the interdependencies between tasks and extracts refined features. It is followed by the quantum-interference based fusion of multi-modalities and the modeling of contextuality. The complex-valued representation, which captures the uncertainty in human language, also plays an important role.

5.7 Error Analysis

We perform an error analysis and show a few misclassification cases (utterance+image), including the cases that MTL predicts correctly while STL fails, and that both setups fail to predict correctly.

From Table 7 and Figure 2, we notice that misclassification for STL often happens in the situation where the literal meaning of an ironic expression differs from its real sentiment. Through utilizing the sentiment knowledge, MTL obtains a significant improvement. Moreover, we observe that MTL might struggle in intricate cases requiring external information.
Table 7: Few error cases where MTL framework performs better than the STL framework.

<table>
<thead>
<tr>
<th>No.</th>
<th>Utterances</th>
<th>Sarcasm (T+V)</th>
<th>Actual</th>
<th>STL</th>
<th>MTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nice job Joe, you are quite the craftsman!</td>
<td>S</td>
<td>N/S</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Ph.D. in electrical engineering made the world laugh without saying a word.</td>
<td>N</td>
<td>N/S</td>
<td>N/S</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Good idea, sit with her. Hold her, comfort her. And if the moment feels right, see if you can cop a feel.</td>
<td>NS</td>
<td>S</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Not a great movie, but look at that beautiful desert.</td>
<td>NS</td>
<td>S</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Those candy canes are making you fatter.</td>
<td>NS</td>
<td>S</td>
<td>S/N</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Wrongly classified visual samples.

Figure 3: Visualization of the commutation relation (a: MUSTARD, b: Memotion) and quantum relative entropy (c: MUSTARD, d: Memotion).

5.8 Discussion on Inter-Task Incompatibility

For a more detailed exploration of the incompatible measurement, we train 1000 and 800 pairs of sentiment and sarcasm measurement operators for MUSTARD and Memotion respectively, and calculate the commutation relation for each pair. The results are visualized in Figure 3a and 3b. We can notice a violation of the commutation law, i.e., $[M_{\gamma}^{sar}, M_{\delta}^{sen}] \neq 0$ for all pairs, implying sentiment and sarcasm are incompatible. To further validate this observation, we introduce quantum relative entropy $^4 D(\sigma||\rho) = Tr\sigma log\sigma - Tr\sigma log\rho$. Here $\sigma$ and $\rho$ are two measurement operators, $Tr$ means the trace operation

Table 8: The correlation between sentiment and sarcasm tasks.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Avg.</th>
<th>Sample Correlation Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUSTARD</td>
<td>0.484</td>
<td>0.517  0.422  0.448  0.461  0.437  0.494</td>
</tr>
<tr>
<td>Memotion</td>
<td>0.461</td>
<td>0.471  0.487  0.677  0.576  0.403  0.401</td>
</tr>
</tbody>
</table>

6 Related Work

(a) Multi-Modal Sarcasm Detection. Schifanella et al. (2016) studied the relationship between textual and visual posts from three major social platforms. Cai et al. (2019) proposed a hierarchical fusion model for multi-modal sarcasm detection. Li et al. (2020) presented an approach based on the state-of-the-art visiolinguistic model ViLBERT. Similarly, Wang et al. (2020) proposed an image-text model for sarcasm detection using the pre-trained BERT and ResNet. Pan et al. (2020) proposed a BERT-based model, which concentrated on both intra and inter-modality incongruity.

(b) Multi-modal Sentiment Analysis. Most recent multi-modal sentiment analysis work is performed from a multi-modal deep learning perspective (Cambria et al., 2019; Kumar and Garg, 2019). Zadeh et al. (2017) introduced a tensor fusion network to fuse audio and visual features. Huang et al. (2019) proposed a deep multi-modal attentive fusion approach. Poria et al. (2019) created the first multi-modal multi-party conversational dataset, namely MELD. Furthermore, Firdaus et al. (2020) and Yu et al. (2020) presented their datasets, i.e., MEISD and CH-SIMS.

Remarkable progress has been made in the current state-of-the-art. However, there is yet lack of mechanisms to capture the inherent uncertainty in multimodal human language for sarcasm and sentiment detection. Different from existing studies, we tackle the problem from a general cognitive perspective with a quantum probabilistic framework.

$^4 D(\sigma||\rho) = Tr\sigma log\sigma - Tr\sigma log\rho$. Here $\sigma$ and $\rho$ are two measurement operators, $Tr$ means the trace operation.
7 Conclusions

We have proposed a quantum probability driven multi-task learning framework. The main idea is to treat each utterance as a complex-valued vector. The contextual interaction between utterances and the correlations across modalities are modeled via quantum composition and quantum interference. Quantum incompatible measurement is performed to yield the probabilistic outcomes. The experimental results verify the effectiveness of the QPM.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (grant No. 2018YFC0831704), Natural Science Foundation of China (grant No. U1636203, U1736103). Also supported by Natural Science Foundation of China (grant No. 62006212), and the fund of State Key Lab. for Novel Software Technology in Nanjing University under grant No.KFKT2021B41, the Doctoral Scientific Research Foundation of Henan Province (grant No. 2020BSJJ031).

We thank for Dr. Qiuchi Li and Benyou Wang’s help.

References


