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

Yaochen Liu[1], **Yazhou Zhang**[2,3*], **Qiuchi Li**[4], **Benyou Wang**[5], **Dawei Song**[1†]
[1]Beijing Institute of Technology
[2]Zhengzhou University of Light Industry
[3] State Key Lab. for Novel Software Technology, Nanjing University
[4]University of Copenhagen, [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

[*]Yazhou Zhang and Yaochen Liu contribute equally and share the co-first authorship.

[†]Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 871–880
November 7–11, 2021. ©2021 Association for Computational Linguistics

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 mutli-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, MUStARD 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 modelling 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 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, $\rho = \sum_i p_i \, |u\rangle \, \langle u|$, 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 path2, 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 path1 and path2. α and β are complex numbers. Its probability is:

$$\begin{aligned} P(x) &= |\varphi_p(x)|^2 = |\alpha\varphi_1(x) + \beta\varphi_2(x)|^2 \\ &= |\alpha\varphi_1(x)|^2 + |\beta\varphi_2(x)|^2 + 2|\alpha\beta\varphi_1(x)\varphi_2(x)| \cos\phi \end{aligned} \tag{1}$$

where ϕ 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 $|u\rangle\langle u|$.

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 $z\left(x\right)$ represents a complex probability amplitude of an event x, where $z\left(x\right) = re^{i\theta}$. QP defines the modulus square of this complex probability amplitude to represent a classical probability $p\left(x\right) = \left|z\left(x\right)\right|^2 = r^2$. 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., $p\left(x = w\right) = \frac{1}{2}$, then the corresponding probability amplitude may be $z\left(x = w\right) = \frac{\sqrt{2}}{2}e^{i\frac{\pi}{4}}$ or $z\left(x = w\right) = \frac{\sqrt{2}}{2}e^{i\frac{3\pi}{5}}$, etc. The amplitude r links to the probability, while the phase θ 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 $z\left(x\right)$ 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:

$$z_3\left(x_3\right) = \alpha z_1\left(x_1\right) + \beta z_2\left(x_2\right) \tag{2}$$

where α and β are complex coefficients. The probabilities of x_1 and x_2 are measured as:

$$p\left(x_1\right) = \left|\alpha\right|^2\left|z_1\left(x_1\right)\right|^2, \;\; p\left(x_2\right) = \left|\beta\right|^2\left|z_2\left(x_2\right)\right|^2 \tag{3}$$

We can derive the probability of multi-modality:

$$\begin{aligned} p\left(x_3\right) &= \left|z_3\left(x_3\right)\right|^2 = \left|\alpha z_1\left(x_1\right) + \beta z_2\left(x_2\right)\right|^2 \\ &= p\left(x_1\right) + p\left(x_2\right) + 2\sqrt{p\left(x_1\right)p\left(x_2\right)}cos\phi \\ &= p\left(x_1\right) + p\left(x_2\right) + \sqrt{p\left(x_1\right)p\left(x_2\right)}\left(e^{i\phi} + e^{-i\phi}\right) \end{aligned} \tag{4}$$

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 ϕ. 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:

$$|u_i\rangle = \alpha_1|w_1\rangle + \beta_1|w_2\rangle, \;\; |u_j\rangle = \alpha_2|w_1\rangle + \beta_2|w_2\rangle \tag{5}$$

The contextual interaction between utterances u_i and u_j constructs the state space of a composite system $\mathcal{H}_{u_i,u_j}$, which is defined as a tensor product of the individual state spaces $|u_i\rangle$ and $|u_j\rangle$:

$$\begin{aligned} \mathcal{H}_{u_i,u_j} &= |u_i\rangle \otimes |u_j\rangle \\ &= \alpha_1\alpha_2|w_1w_1\rangle + \alpha_1\beta_2|w_1w_2\rangle \\ &+ \beta_1\alpha_2|w_2w_1\rangle + \beta_1\beta_2|w_2w_2\rangle \end{aligned} \tag{6}$$

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, $M^{sar} = \left\{M_{\gamma}^{sar}\right\}_{\gamma=1}^{G}$, $M^{sen} = \left\{M_{\delta}^{sen}\right\}_{\delta=1}^{G}$. If any cross-task pair of measurement operators satisfy the commutation rule[1], i.e., $\left[M_{\gamma}^{sar}, M_{\delta}^{sen}\right] = 0$ for all γ and δ, 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] $\left[M_{\gamma}^{sar}, M_{\delta}^{sen}\right] = M_{\gamma}^{sar}M_{\delta}^{sen} - M_{\delta}^{sen}M_{\gamma}^{sar} = 0$

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 multi-modal samples. The ξ^{th} sample X^{ξ} is represented as $\left\{X^{\xi} = \left(C^{i}, U^{\xi}\right), Y^{\xi}\right\}$, where C^{i}, U^{ξ}, Y^{ξ} denote the i^{th} conversational context, the multi-modal utterance and the label respectively, and $i \in [1, 2, ..., k]$, $\xi \in [1, 2, ..., L]$. Both the context and the multi-modal utterance consist of textual and visual modalities, i.e., $C^{i} = \left(C_{t}^{i}, C_{v}^{i}\right)$, $U^{\xi} = \left(U_{t}^{\xi}, U_{v}^{\xi}\right)$.

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

$$\zeta = \prod_{i} p\left(Y^{\xi} | C^{i}, U^{\xi}, \Theta\right) \tag{7}$$

where Θ represents the parameter set.

Overall Network. The overall architecture of the QPM framework is shown in Figure 1. (1) The ξ^{th} textual utterance and its visual counterpart are represented by complex-valued embeddings, denoted as $|u_{t}^{\xi}\rangle$ and $|u_{v}^{\xi}\rangle$. (2) Then, $|u_{t}^{\xi}\rangle$ and $|u_{v}^{\xi}\rangle$ are fed into the quantum composition layer to capture the contextuality, where the results are encapsulated in two density matrices ρ_{text} and ρ_{img}. (3) We then fuse ρ_{text} and ρ_{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 $|\{w_{t}^{j}\rangle\}_{j=1}^{n}$. With words as the basic semantic unit, the j^{th} word w_{t}^{j} can be used as the basis state $|w_{t}^{j}\rangle$, represented by one-hot encoding, i.e., the j-th element being 1 and 0s elsewhere.

Then, we regard the ξ^{th} target utterance u_{t}^{ξ} as a quantum superposition of a set of basis words $\left\{|w_{t}^{1}\rangle, |w_{t}^{2}\rangle, ..., |w_{t}^{n}\rangle\right\}$, which is formulated as:

$$|u_{t}^{\xi}\rangle = \sum_{j=1}^{n} z_{t}^{j} |w_{t}^{j}\rangle, \quad z_{t}^{j} = r_{t}^{j} e^{i\theta_{t}^{j}} \tag{8}$$

where n is the number of words in the utterance. z_{t}^{j} is a complex probability amplitude expressed in the polar form. i is the imaginary number. r_{t}^{j} is the modulus of the complex number, termed amplitude. $\theta_{t}^{j} \in [-\pi, \pi]$ is the argument (phase) of z_{t}^{j}.

We construct the complex-valued vector of the ξ^{th} utterance, by associating the amplitude r with the semantic knowledge and the phase θ with the pre-assigned sentiment orientation, i.e., $|u_{t}^{\xi}\rangle = \left(r_{t}^{1} e^{i\theta_{t}^{1}}, r_{t}^{2} e^{i\theta_{t}^{2}}, ..., r_{t}^{n} e^{i\theta_{t}^{n}}\right)^{T}$.

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 $\{|w_{v}^{j}\rangle\}_{j=1}^{n}$, where the visual part of the target utterance is represented as $|u_{v}^{\xi}\rangle$.

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 $|u_{t}^{\xi}\rangle$ and its contexts $\left\{|c_{t}^{1}\rangle \ldots |c_{t}^{k}\rangle\right\}$, the contextual interaction between them constructs a textual composite system $\Psi_{t}^{\xi,k}$, 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 λ^{th} composite system is computed as:

$$|\Psi_{t}^{\xi,\lambda}\rangle| = |u_{t}^{\xi}\rangle \otimes |c_{t}^{1}\rangle \otimes |c_{t}^{2}\rangle \otimes, ..., \otimes |c_{t}^{\lambda}\rangle \tag{9}$$

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

These k composite systems are mathematically encapsulated in a textual density matrix ρ_{text}, to obtain the representation of the target utterance u_{t}^{ξ}.

$$\rho_{text} = \sum_{\lambda=1}^{k} p_{\lambda} |\Psi_{t}^{\xi,\lambda}\rangle\langle\Psi_{t}^{\xi,\lambda}| \tag{10}$$

where p_{λ} 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., $\Psi_{v,k} =$

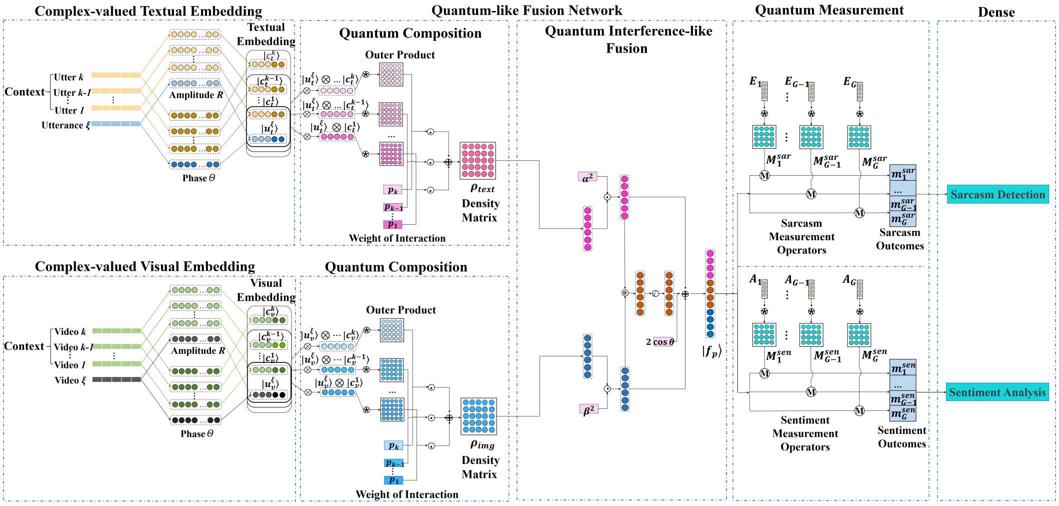


Figure 1: The architecture of the QPM framework. $\otimes$ denotes the tensor product operation. $\circledast$ denotes an outer production to a vector. $\odot$ denotes point-wise multiplication. $\oplus$ refers to a element-wise addition. $\circledcirc$ is the matrix multiplication. √⃝ refers to the square operation. Ⓜ refers to the quantum measurement operation.

$\left\{|\Psi_v^{\xi,1}\rangle, |\Psi_v^{\xi,2}\rangle, ..., |\Psi_v^{\xi,k}\rangle\right\}$, and obtain the visual density matrix ρ_{img} using Eq. 10.

Then, textual and visual density matrices ρ_{text} and ρ_{img} are flattened into two vectors $|f_t\rangle$ and $|f_v\rangle$ 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:

$$z_p(x) = \alpha z_t(x) + \beta z_v(x) \tag{11}$$

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

$$f_p(x) = f_t(x) + f_v(x) + 2\sqrt{f_t(x)f_v(x)}\cos\phi_i \tag{12}$$

where x is the x^{th} feature component of the multi-modal representation $|f_p\rangle$. $I = 2\sqrt{f_t(x_i)f_v(x_i)}\cos\phi_i$ is the interference item. $|f_p\rangle = (f_p(x_1), f_p(x_2), ..., f_p(x_n))^T$ 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 $\vec{m}^{sar}$ and $\vec{m}^{sen}$.

Specifically, two sets of measurement operators $M^{sar} = \{M_\gamma^{sar}\}_{\gamma=1}^G$, $M^{sen} = \{M_\delta^{sen}\}_{\delta=1}^G$ are pre-defined, each constructed by the outer product of the corresponding measurement vector $|E_\gamma\rangle$ or $|A_\delta\rangle$, i.e., $M_\gamma^{sar} = |E_\gamma\rangle\langle E_\gamma|$, $M_\delta^{sen} = |A_\delta\rangle\langle A_\delta|$. The probability distribution over the measurement outcomes can be computed as: $\vec{m}^s = tr\left((M^s)^\dagger M^s|f_p\rangle\langle f_p|\right)$, where $s \in \{sar, sen\}$.

4.6 Dense Layer

The sarcastic and sentimental outcomes $\vec{m}^{sar}$, $\vec{m}^{sen}$ 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 ζ_{sar} and ζ_{sen}, and jointly minimize them with different weights, e.g., $\zeta = w_{sar}\zeta_{sar} + w_{sen}\zeta_{sen}$. 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

[2] http://www.iitp.ac.in/ai-nlp-ml/resources.html

Dataset	*Task*	*Classes*	*No. of Utter.*	*RC(%)*
Memotion	Sarcasm	Sar.	5448	77.92
		Non.	1544	22.08
	Sentiment	Pos.	631	9.02
		Neg.	4160	59.50
		Neu.	2201	31.48
MUStARD$_{ext}$	Sarcasm	Sar.	345	50.00
		Non.	345	50.00
	Sentiment	Pos.	210	30.43
		Neg.	391	56.67
		Neu.	89	12.90

Table 1: Dataset statistics.

Hyper-parameters	*MUStARD$_{ext}$*	*Memotion*
Embedding size	768	
Activations	Relu	
Batch	48	
Learning rate	0.001	0.003
No. of measurement	1000	800
dropout	0.6	0.5
Interference item $\cos\phi_i$	-0.3	0.2
(α^2, β^2)	(0.7,0.3)	(0.8,0.2)

Table 2: Model configurations.

Dataset	**Method**	**Sarcasm Detection**		
		P	R	M$_i$-F1
MUStARD$_{ext}$	SVM+BERT	65.14	64.61	64.68
	SVM+BERT (+context)	65.53	65.11	65.06
	RCNN-RoBERTa	68.70	64.33	65.16
	EfficientNet	63.58	64.19	63.77
	UPB-MTL	65.12	65.41	65.41
	QMSA	70.23	70.04	70.00
	A-MTL	77.09	76.67	76.57
	Text-QPM	72.07	72.34	72.12
	Image-QPM	65.36	65.46	65.42
	QPM	**77.49**	**77.61**	**77.53**
	△SOTA	(+0.5%)	(+1.3%)	(+1.3%)
Memotion	SVM+BERT	44.17	44.36	44.15
	SVM+BERT (+context)	45.11	45.22	45.04
	RCNN-RoBERTa	50.44	50.77	50.52
	EfficientNet	50.59	50.81	50.75
	UPB-MTL	51.38	51.71	51.59
	QMSA	55.84	56.36	56.42
	A-MTL	60.23	59.74	59.85
	Text-QPM	51.29	51.05	51.12
	Image-QPM	51.69	51.87	51.87
	QPM	**61.42**	**61.07**	**61.39**
	△SOTA	(+2.0%)	(+2.2%)	(+2.1%)

Table 3: Comparison of different models.

et al., 2020) and Memotion[3] (Sharma et al., 2020) datasets meet these criteria. **MUStARD$_{ext}$:** The utterance in each dialogue is annotated with sarcastic or non-sarcastic labels. As an extended version of MUStARD, MUStARD$_{ext}$ re-annotate sentiment and emotion labels. **Memotion:** It consists of 6992 training samples and 1879 testing samples. Each memo data has been labelled with semantic dimensions, e.g., sentiment, sarcasm, humor, etc. Table 1 shows the detailed statistics for these two datasets.

Evaluation metrics. We adopt *precision* (P), *recall* (R) and *micro-F1* (M$_i$-F1) as evaluation metrics in our experiments. We also introduce a *balanced accuracy* metric for an ablation test.

Hyper-parameter Setup. The textual and visual amplitudes are initialized with BERT and ResNet152 respectively. The phases are initialized with the pre-assigned sentiments using BERT. The quantum measurements are randomly initialized with an unit vector and is set to be trainable. The optimal hyper-parameters are listed in Table 2.

5.2 Baselines

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 both quantum-like fusion network and quantum measurement mechanism, which ensures that QPM can model intra-modality contextuality and inter-

[3] https://competitions.codalab.org/competitions/20629

Task Dataset	**Setups**	**T**		**V**		**T+V**	
		M$_\text{i}$-F1	Acc	M$_\text{i}$-F1	Acc	M$_\text{i}$-F1	Acc
Sarcasm MUStARD	STL	62.51	62.48	64.00	64.00	66.37	66.21
	MTL	72.12	72.04	65.42	65.34	77.53	77.50
Sentiment MUStARD	STL	55.31	55.19	57.54	57.50	60.00	60.00
	MTL	55.43	55.40	62.36	62.14	66.11	66.05
Sarcasm Memotion	STL	50.47	50.50	51.62	51.62	52.11	52.03
	MTL	51.12	51.07	51.87	52.04	61.39	61.45
Sentiment Memotion	STL	37.54	37.60	37.33	37.42	39.23	39.14
	MTL	42.10	41.22	41.26	41.26	42.67	42.70

Table 4: Comparison with single-task learning (STL) and multi-task (MTL) learning frameworks. T: Text, V: Visual, T+V: QPM

modality interference, and refine the final features.

5.4 STL v/s MTL Framework

We outline the comparison results between the multi-task (MTL) and single-task (STL) learning frameworks in Table 4. Bi-modal (T+V) shows a better performance over unimodal setups.

For sarcasm detection, MTL outperforms STL by a large margin in text modality and bi-modal. The reason is that visual sarcasm detection involves a higher level of abstraction and more subjectivity. For sentiment analysis, MTL with sarcasm together achieves better performance than STL on all modalities. This indicates that sarcasm assists sentiment analysis through the sharing of knowledge, and vice versa. Our QP-based MTL framework could learn the inter-dependence between two related tasks and improves performance.

5.5 Effect of Context Range

Since the Memotion dataset does not involve contexts, we only report results on MUStARD$_{ext}$ in Tables 5 with different context scopes. “Zero context” means that we only use the target utterance, ignoring its context. “One context” denotes that we use one previous utterance to construct the density matrix. “Two contexts” means the use of previous two utterances as context.

The performance steadily increases as context range increases (with F1 scores of 66.03%, 68.75%, 72.54% and 77.53%), showing the importance of incorporating conversational context. QPM with zero context unsurprisingly performs worst. QPM with all contexts achieves the best F1 score, implying that incorporating all conversational contexts would be the best way to reach an optimal performance.

5.6 Ablation Study

We perform an ablation study to further study the effectiveness of different components of QPM: (1) *QPM-Real* 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 repalces the quantum interference-like fusion layer with multi-modal concatenation; (4) *QPM-Trad* that replaces quantum incompatible measurements with traditional softmax layers.

Dataset	**Context range**	**Metrics**	
		M$_\text{i}$-F1	Acc
MUStARD$_{ext}$	Zero	66.03	66.03
	One	68.75	68.67
	Two	72.54	72.47
	All	77.53	77.50

Table 5: Effect of context range.

Dataset	**Models**	**Metrics**	
		M$_\text{i}$-F1	Acc
MUStARD$_{ext}$	QPM-Real	70.03	70.01
	QPM-Speaker Independent	66.22	66.09
	QPM-Concat	66.13	66.04
	QPM-Trad	62.31	62.18
	QPM	77.53	77.50
Memotion (No context)	QPM-Real	53.48	53.48
	QPM-Concat	52.64	52.64
	QPM-Trad	52.08	52.11
	QPM	61.39	61.45

Table 6: Ablation experiment results.

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 modelling 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.

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

Table 7: Few error cases where MTL framework performs better than the STL framework.

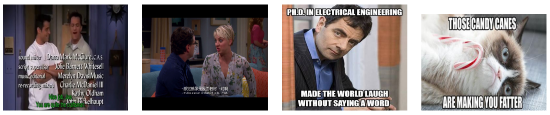
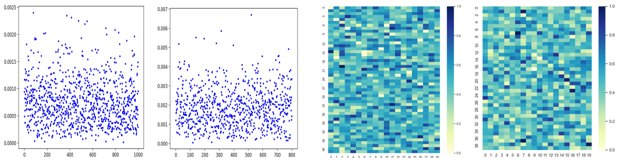

PH.D. IN ELECTRICAL ENGINEERING
MADE THE WORLD LAUGH WITHOUT SAYING A WORD
THOSE CANDY CANES
ARE MAKING YOU FATTER

Figure 2: Wrongly classified visual samples.



(a) (b) (c) (d)

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., $\left[M_{\gamma}^{sar}, M_{\delta}^{sen}\right] \neq 0$ for all pairs, implying sentiment and sarcasm are incompatible. To further validate this observation, we introduce quantum relative entropy[4], which is a kind of "distance" measure between quantum states, the smaller quantum relative entropy show the closer correlation between sentiment and sarcasm operators. Average correlation and sample correlation scores are presented in Table 8 and Figure 3c, 3d, showing the two tasks are correlated. The result justifies the need of incompatible measurement and explains its effectiveness against traditional multi-task learning setting in Table 6.

Furthermore, an analysis of data shows that 84% of sarcasm samples in MUStARD express explicit sentiments while the proportion in Memotion is 74%. In MUStARD 38% of ironic utterances are also positive, and in Memotion it is 36%. These results support our hypothesis that sarcasm and sentiment are closely related.

Dataset	Avg.	Sample Correlation Scores					
MUStARD	0.484	0.517	0.422	0.448	0.461	0.437	0.494
Memotion	0.461	0.471	0.487	0.677	0.576	0.403	0.401

Table 8: The correlation between sentiment and sarcasm tasks.

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 σ and ρ 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

Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. 2009. Is there something quantum-like about the human mental lexicon? *Journal of mathematical psychology*, 53(5):362–377.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Erik Cambria, Soujanya Poria, and Amir Hussain. 2019. Speaker-independent multimodal sentiment analysis for big data. In *Multimodal Analytics for Next-Generation Big Data Technologies and Applications*, pages 13–43. Springer.

Noël Carroll and Noël E Carroll. 1999. *Philosophy of art: A contemporary introduction*. Psychology Press.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. 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*, pages 4351–4360, Online. Association for Computational Linguistics.

Sébastien Designolle, Máté Farkas, and Jędrzej Kaniewski. 2019. Incompatibility robustness of quantum measurements: a unified framework. *New Journal of Physics*, 21(11):113053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dimitris Gkoumas, Qiuchi Li, Shahram Dehdashti, Massimo Melucci, Yijun Yu, and Dawei Song. 2021. Quantum cognitively motivated decision fusion for video sentiment analysis. *arXiv preprint arXiv:2101.04406*.

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37.

Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. 2020. A quantum interference inspired neural matching model for ad-hoc retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–28.

Akshi Kumar and Geetanjali Garg. 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, pages 1–17.

Lily Li, Or Levi, Pedram Hosseini, and David A Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. *arXiv preprint arXiv:2010.06671*.

Qiuchi Li, Dimitris Gkoumas, Christina Lioma, and Massimo Melucci. 2021. Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, 65:58–71.

Qiuchi Li, Benyou Wang, and Massimo Melucci. 2019. Cnm: An interpretable complex-valued network for matching. In *NAACL-HLT (1)*, pages 4139–4148.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1383–1392.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 527–536.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, pages 1–12.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 1136–1145, New York, NY, USA. Association for Computing Machinery.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

Alessandro Sordoni, Jian-Yun Nie, and Yoshua Bengio. 2013. Modeling term dependencies with quantum language models for ir. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 653–662, New York, NY, USA. Association for Computing Machinery.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114.

Sagar Uprety, Dimitris Gkoumas, and Dawei Song. 2020. A survey of quantum theory inspired approaches to information retrieval. *ACM Computing Surveys (CSUR)*, 53(5):1–39.

George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin-Gabriel Chiru, and Stefan Trausan-Matu. 2020. Upb at semeval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1208–1214.

Benyou Wang, Qiuchi Li, Massimo Melucci, and Dawei Song. 2019. Semantic hilbert space for text representation learning. In *The World Wide Web Conference on*, pages 3293–3299.

Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. 2020. 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*, pages 19–29.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. 2018a. End-to-end quantum-like language models with application to question answering. pages 5666–5673.

Peng Zhang, Zhan Su, Lipeng Zhang, Benyou Wang, and Dawei Song. 2018b. A quantum many-body wave function inspired language modeling approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1303–1312.

Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. 2020. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 62:14–31.

Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. 2018c. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science*, 752:21–40.