The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

# Quantum-inspired Neural Network for Conversational Emotion Recognition

**Qiuchi Li[1], Dimitris Gkoumas[2], Alessandro Sordoni[3], Jian-Yun Nie[4], Massimo Melucci[1]**
[1] University of Padua, Padua, Italy
[2] The Open University, Milton Keynes, UK
[3] Microsoft Research Montreal, Montreal, Canada
[4] Université de Montréal, Montreal, Canada
[1]{qiuchili,melo}@dei.unipd.it, [2]dimitris.gkoumas@open.ac.uk, [3]alsordon@microsoft.com, [4]nie@iro.umontreal.ca

## Abstract

We provide a novel perspective on conversational emotion recognition by drawing an analogy between the task and a complete span of quantum measurement. We characterize different steps of quantum measurement in the process of recognizing speakers' emotions in conversation, and stitch them up with a quantum-like neural network. The quantum-like layers are implemented by complex-valued operations to ensure an authentic adoption of quantum concepts, which naturally enables conversational context modeling and multimodal fusion. We borrow an existing algorithm to learn the complex-valued network weights, so that the quantum-like procedure is conducted in a data-driven manner. Our model is comparable to state-of-the-art approaches on two benchmarking datasets, and provide a quantum view to understand conversational emotion recognition.

## Introduction

Multimodal conversational emotion recognition is a new but rapid-growing area. The task is to classify each utterance in a conversation into one of the candidate emotions based on clues from multimodal channels. A speaker's emotion is expressed not only by words, but also from his facial emotions and speech voices. The recognition of emotion in a conversation hence requires a joint analysis of multimodal data including textual, visual and acoustic modalities. Figure 1 is an example of a multimodal conversation between three speakers (parties), Joey, Monica and Phoebe. The emotions of all speakers dramatically change in the course of conversation. Hence, we are facing with a challenge of automatically tracking the emotion evolution.

Existing works have mainly managed to model two levels of interaction. On the one hand, unimodal features are merged into a joint multimodal utterance representation, in which interactions between different modalities are captured (i.e. multimodal fusion) (Liang et al. 2018; Zadeh et al. 2018a; Tsai et al. 2019a; Zadeh et al. 2017; Zhang et al. 2020). On the other hand, the speakers' interactions in a conversation are captured based on RNN-based backbone structures (i.e. conversational context modeling) (Poria et al. 2017; Hazarika et al. 2018a,b; Ghosal et al. 2019; Majumder et al. 2019; Zhang et al. 2020). However, few works



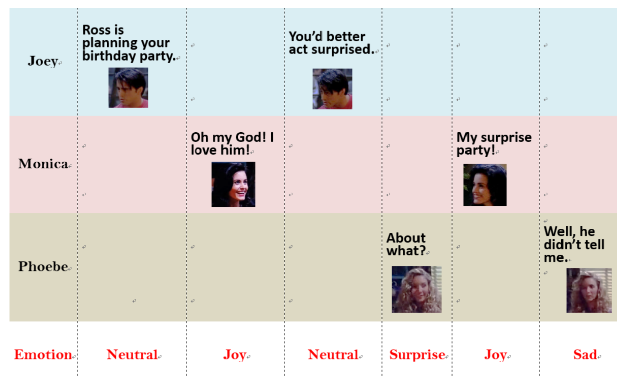


Figure 1: An example of a multimodal conversation. The task is to predict the emotion of each utterance.

have joined multimodal fusion and conversational context modeling in a unified architecture. Most multimodal fusion works are evaluated on monologue data with no conversational structure involved. For conversational context modeling, however, simple concatenation or attention mechanism is employed to join pre-trained unimodal utterance features. Another issue facing both aspects of research is a lack of formal understanding of the constructed model, which is mainly composed of black-box-like neural components (Baltrušaitis, Ahuja, and Morency 2017).

We design a quantum-like framework to approach conversational emotion recognition, which tackles both limitations in one shot. The motivation stems theoretical investigations of quantum cognition (Busemeyer and Bruza 2012), which suggest that quantum-inspired frameworks can properly explain phenomena in human cognition that violate the probability theory that grounds almost all classical models. As a typical cognitive concept, emotion recognition has received little attention from a quantum viewpoint. We therefore seek to explore the use of quantum-like procedure to model emotion recognition.

We draw an analogy between the process of quantum measurement and the emotion recognition. In a quantum physics experiment, a particle is in a *mixture* of multiple mutually independent *pure states* prior to measurement, and the measurement makes it collapse onto a single pure measurement state. Likewise, a speaker is in an ambiguous state of multiple independent emotions, and the conversational context serves as a measurement that causes the emotion state to collapse onto the a pure state. Moreover, the evolution of quantum states over time is analogous to the evolution of a speaker's emotion state in the course of conversation.

This analogy stimulates us to contrive the procedure of a quantum measurement experiment for conversational emotion recognition (Ringbauer 2017). As complex values are key to instrument quantum concepts, we build a complex-valued neural network to implement this measurement procedure. In addition, a dedicated optimizer (Wisdom et al. 2016) is employed to update the complex-valued unitary matrices manifested in the representation of quantum concepts, so that the whole model can be trained end-to-end with standard back-propagation algorithms. This allows us to determine the specifications of the pre-designed quantum-like process in a data-driven manner.

We evaluate our framework on two benchmarking conversational emotion recognition datasets, namely MELD (Poria et al. 2019a) and IEMOCAP (Busso et al. 2008a). The results show that the provided formal quantum view of conversational emotion recognition does not lead to the drop in performance: our model achieves comparable accuracy performances to state-of-the-art models on both datasets, with slightly improved values on particular metrics. Moreover, the introduced training algorithm for unitary matrix brings to affordable drop in efficiency.

Our contributions are as follows:

- We take a novel quantum perspective on conversational emotion recognition.
- We build a unified framework to simultaneously conduct multimodal fusion and conversational context modeling.
- We design a set of complex-valued network layers to implement the quantum concepts, involving unitary matrices. We manage to make the neural network end-to-end trainable.
- We conduct a comprehensive and fair comparison with existing models, and our model achieves comparable performances to the state-of-the-art model.

## Preliminaries on Quantum Theory

Quantum physics (QT) provides a mathematical interpretation of the microscopic world such as electrons and photons. The mathematical formalism of quantum physics is defined on a *Hilbert Space* $\mathcal{H}$, which is an inner product space over the complex field. We employ the widely-used *Dirac Notations* for a mathematical representation of quantum concepts. A complex-valued *unit* vector $\vec{\mu}$ and its conjugate transpose $\vec{\mu}^H$ are denoted as a *ket* $|u\rangle$ and a *bra* $\langle u|$ respectively[1].

### State

The state of an isolated quantum system is called a *quantum state*, such as the position or momentum of an electron. If the system composed solely of a single particle, its state is then a *pure state* $|\phi\rangle$, which is a unit complex vector on $\mathcal{H}$. In particular, when the pure state falls onto a basis of the Hilbert Space, it is called a *basis state*. Otherwise, it is a *superposition* of the basis states $|0\rangle$ and $|1\rangle$ and called a *superposition state*.

When the quantum system is composed of multiple particles, the overall system state is a statistic *mixture* of individual particle states or a *mixed state*. A mixed state is mathematically a *density matrix*, which is a positive semi-definite square matrix with unit trace. For the set of pure states $\{|\phi_j\rangle\}_{j=1}^n$ with weights $\{p_j\}_{j=1}^n$ that sum up to 1, the density matrix $\rho$ is computed by $\rho = \sum_{i=1}^n p_j |\phi_j\rangle \langle\phi_j|$. It is worth noting that density matrix can be viewed as a generic state representation, since a pure state $|\phi_k\rangle$ can be recast to a density matrix via $\rho = |\phi_k\rangle \langle\phi_k|$.

## A Complete Procedure of Measurement Experiment

*Measurement* is the process of measuring the physical property of a system. A complete span of quantum measurement in a lab experiment contains state *preparation*, *evolution*, *measurement* and *collapse*. Below are brief introductions of all steps. For details, please refer to Chapter 2 in (Ringbauer 2017).

**Preparation** State preparation is literally the process of preparing the quantum system. After this process, the state $\rho$ of the system to be measured is obtained.

**Evolution** The prepared system does not remain unchanged, but undertakes a complicated evolution process over time before the measurement. The evolution can be mathematically formulated as a *Unitary Operator* or equivalently a complex unitary matrix $U \in \mathcal{H}$, satisfying $UU^H = I$[2]. The evolution makes the state change to

$$\hat{\rho} = U\rho U^H \tag{1}$$

It is worth noting that result $\hat{\rho}$ is also a density matrix as long as the input $\rho$ is a density matrix. So a valid physical state is produced after the evolution step.

**Measurement** A measurement is associated to an *observable* $\hat{O}$, which is a self-joint square matrix in the Hilbert Space, i.e. $\hat{O} = \hat{O}^H$[3]. An observable can be eigen-decomposed into

$$\hat{O} = \sum_j \lambda_j |\lambda_j\rangle \langle\lambda_j| \tag{2}$$

where the eigenstates $\{|\lambda_j\rangle\}$ form a complete orthogonal basis of the Hilbert Space $\mathcal{H}$ while the eigenvalues $\{\lambda_j\}$ are the possible observed values. For a system $\rho$, the probability $p_j$ that $\lambda_j$ is observed is given by the Born's rule (Born 1926):

[1]Hence The inner and outer product of two unit vectors $|u\rangle$ and $|v\rangle$ are $\langle u|v\rangle$ and $|u\rangle \langle v|$ respectively.

[2]$A^H$ is called the Hermitian of matrix A, meaning the conjugate transpose

[3]Here a nomenclature rather than a strict definition is used for understanding purpose. Please refer to (Nielsen and Chuang 2011) for a strict definition of projection measurement.

$$p_j = tr(\hat{\rho}\,|\lambda_j\rangle\,\langle\lambda_j|) = \langle\lambda_j|\,\hat{\rho}\,|\lambda_j\rangle \quad (3)$$

the resulting probabilities $\{p_j\}$ form a classical probability distribution with $\sum p_j = 1$.

**Collapse** After measurement, the system is always collapsed onto one pure eigenstate $|\lambda_k\rangle$ of the observable. If the measurement can be repeated for infinite times, then at probability $p_k$ (computed by Eq.3) the system collapses onto state $|\lambda_k\rangle$.

# Related Works

## Multimodal Fusion

Multimodal fusion approaches are targeted at monologue data, mainly based on word-aligned multimodal features. Beyond simple concatenation of features under recurrent structures, hybrid memories have been constructed by introducing an additional cell that aggregates the hidden units of unimodal recurrent structures at a time stamp, and is fed to the next time stamp as an additional input (Liang et al. 2018; Zadeh et al. 2018c,a; Bagher Zadeh et al. 2018). Sequence-to-sequence structures have also been employed to "translate" one modality representation to another for the same utterance, and take the hidden representation as the joint utterance representation (Pham et al. 2019; Tsai et al. 2019a). Other models rely on tensor-based approaches to fuse multimodal features, considering the natural split in terms of the modalities (Zadeh et al. 2017; Liu et al. 2018; Barezi and Fung 2019; Liang et al. 2019; Mai, Hu, and Xing 2019) to form a tensorized representation for a multimodal utterance, followed by fully connected network (Zadeh et al. 2017) or tensor decomposition strategies (Liu et al. 2018; Barezi and Fung 2019) to conduct classification.

## Conversational Context Modeling

The works on conversational context modeling are targeted for conversational emotion recognition, either in a textual or multimodal setting. They mainly employ pretrained utterance-level unimodal representations and conduct simple concatenation or attention to obtain utterance representation. One idea is to build a memory cell for each speaker in an attempt to achieve speaker-specific context modeling (Hazarika et al. 2018b,a). However, it is later argued that memory cell does not well exploit the speaker information (Baltrušaitis, Ahuja, and Morency 2017). More recent models (Majumder et al. 2019; Ghosal et al. 2019) replace the memory cell with components to handle self and inter-speaker emotional influence. In particular, DialogueRNN (Majumder et al. 2019) builds a hierarchical multi-stage RNN with different strategies for updating a speaker and a listener's emotion states. DialogueGCN (Ghosal et al. 2019) captures the relations of all utterances in a conversation, based on their relative order and whether they belong to the same speaker. The relations are reflected in a graph, and a graph neural network is built to update utterance representations.

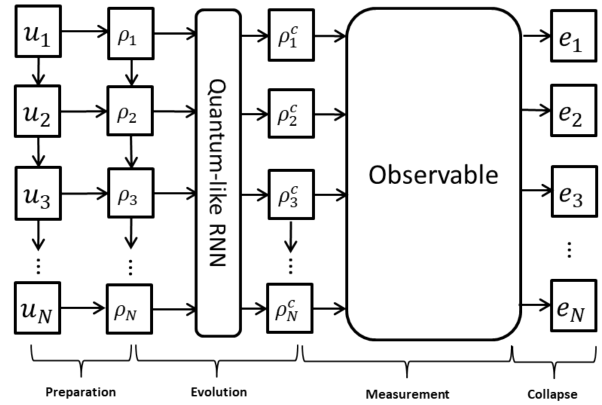


Figure 2: Diagram of the proposed network. Each utterance $u_j$ is a represented by density matrix $\rho_j$. The evolution step is a Quantum-like RNN. Post-evolution states $\{\rho_j^c\}$ are fed into the measurement controlled by observable $O$. The emotions with the largest likelihood $\{e_j\}$ are produced.

# Methodology

## Problem Definition

The task input is a multimodal conversation $S$ containing $N$ utterances $\{u_j\}_{i=1}^N$. Each utterance $u_j$ has textual, visual and acoustic representations $t_j, v_j, a_j$, and uttered by party $p_j$. Suppose there are a total number of $K$ parties in the whole dataset, then $p_j \in \{1, 2, ..., K\}$. The task requires one to predict the emotion $e_j$ for each utterance $u_j$ within a finite set of emotions $E$.

## Unimodal Feature Extraction

We build different neural network structures to extract textual, visual and acoustic features respectively. For textual features, CNN (Kim 2014) is employed to extract textual features from the transcripts, with a 300-dim Glove vector (Pennington, Socher, and Manning 2014) for each word. 3D-CNN (Ji et al. 2013) and openSMILE (Eyben et al. 2010) are utilized to extract the features respectively. Please refer to (Hazarika et al. 2018b) for details of the network structures.

## Quantum-inspired Neural Network for Emotion Recognition in Conversation

Figure 3 shows our model for conversational emotion recognition, termed as Quantum Measurement-inspired Neural Network (QMNN), which consists of four steps, namely *preparation*, *evolution*, *measurement* and *collapse* in correspondence to the quantum measurement procedure.

**Preparation** We prepare the state $\rho_j$ of each utterance $u_j$ in a conversation. A multimodal fusion is conducted by means of quantum mixture. As shown in figure 3, the unimodal features are recast as pure states, and the utterance is viewed as a mixture of the unimodal states.

For the construction of unimodal pure states, we consider the *phase-amplitude or polar decomposition* of a complex value $z$ as $z = re^{i\theta}$, where amplitude $r$ is a non-negative

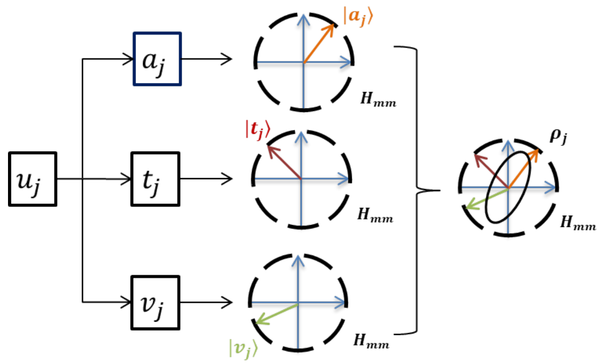


Figure 3: Diagram of preparation. Unimodal states $|a_j\rangle, |t_j\rangle, |v_j\rangle$ are constructed and mixed to produce the multimodal mixed state $\rho_j$.

value, phase $\theta$ is a real value in $[0, 2\pi)$, and $i$ is the imaginary number with $i^2 = -1$. A pure state $|\psi\rangle$ can generally be expressed as

$$\begin{aligned} |\psi\rangle &= [r_1 e^{i\theta_1}, ..., r_d e^{i\theta_d}] \\ &= [r_1, ..., r_d] \odot e^{i[\theta_1, ..., \theta_d]} \end{aligned} \quad (4)$$

Where $\odot$ refers to element-wise vector product. The amplitudes $R = [r_1, ..., r_d]$ forms a real unit vector, while the phases $\Theta = [\theta_1, ..., \theta_d]$ are real vectors with all elements in $[0, 2\pi]$. They are constructed respectively in the formation of unimodal pure states. Suppose the input features are $a_i \in \mathcal{R}^{d_a}, v_i \in \mathcal{R}^{d_v}, t_i \in \mathcal{R}^{d_t}$ for acoustic, visual and textual modalities for utterance $u_i$. The features are first projected to the same $d$-dimensional multimodal Hilbert Space $\mathcal{H}_{mm}$ by a single fully connected layer with Rectified Linear Unit (ReLU) as the activation function: $\hat{m_j} = ReLU(W_m m_j + b_m), m \in \{a, v, t\}$. Then the $d$-dimensional vectors are normalized to produce unimodal pure states: $R_{m_j} = \frac{\hat{m_j}}{||\hat{m_j}||_2}, m \in \{a, v, t\}$. The ReLU function ensures all elements of the normalized vector are non-negative, and the normalized vector can be taken as amplitudes of a pure state.

The phase assignment is motivated by the prior work of encoding position in complex word embeddings (Wang et al. 2020), which demonstrates that complex embedding with periodic phases is one and the only that preserves the relative word distance. Likewise, we encode *utterance order* and *speaker information* in the phases. The phase vector for the $j$-th utterance is calculated by $\Theta_j = W_{p_j} j + \Psi_{p_j}$, where $W_{p_j} \in \mathcal{R}^d$ is the frequency of speaker $p_j$ and $\Psi_{p_j} \in \mathcal{R}^d$ are speaker-dependent initial phases with all elements in $[-\pi, \pi]$. For each modality, $K$ different $d$-dimensional frequency and initial phase vectors are learned from the data. We expect them to capture certain speaker-dependent features such as utterance frequency or emotion tendency in each modality.

Based on the phase encoding $\Theta_j^{(a)}, \Theta_j^{(v)}, \Theta_j^{(t)}$, the unimodal states are constructed by $|m_j\rangle = R_{m_j} \odot e^{i\Theta_j^{(m)}}, m \in \{a, v, t\}$. A mixture process is then in place to fuse the unimodal pure states.

$$\rho_j = \lambda_a |a_j\rangle \langle a_j| + \lambda_v |v_j\rangle \langle v_j| + \lambda_t |t_j\rangle \langle t_j| \quad (5)$$

where $\lambda_a, \lambda_v, \lambda_t$ are non-negative values that sum up to 1. We take the lengths of the projected vectors to compute the mixture weights: $\lambda_a, \lambda_v, \lambda_t = softmax(||\hat{a_j}||_2, ||\hat{v_j}||_2, ||\hat{t_j}||_2)$. This is because the length of the vector is thrown away in the construction of pure states, but it may still contain useful information to the task. In addition, then length of the vector is analogous to the quantities of particles and somehow reflects the mixture weights.

The construction of utterance state naturally entails multimodal fusion and encodes speaker information. During training, different mixture weights are produced for different utterances in a conversation, formulating the evolving influences of each modalities to the final emotion. Encoding speaker information in the phases allows for a complicated non-linear interaction between the speaker features and the multimodal features. The utterance representation gives rise to a latent speaker interaction in the subsequent network architectures.

**Evolution** In the conversational emotion recognition task, the emotions of speakers are evolving throughout the conversation. Hence it is intuitive to employ quantum evolution to track the dynamics of emotional states in a conversation.

The building block of the evolution step is a quantum-like recurrent neural network (QRNN). The inputs of a QRNN is a sequence of quantum states represented by density matrices $\rho_x^1, ..., \rho_x^N$ with $N$ being the sequence length. A hidden density matrix $\rho_h$ is introduced to memorize the sequential information. Its value $\rho_h^t$ at time $t$ is iteratively updated by

$$\rho_h^t = \lambda U_h \rho_h^{t-1} U_h^* + (1-\lambda) U_x \rho_x^t . U_x^*, t = 1, ..., N \quad (6)$$

It is easy to check that the result matrix $\rho_h^t$ is still a legal density matrix. With initial value of density matrix $\rho_h^0$ being a random diagonal matrix with unit trace, $U_x, U_h$ are complex-valued unitary matrices, and $\lambda \in [0, 1]$. From a quantum language, the process means the state of the context is evolving over time, and mixing with the input state at each time stamp.

Similar to classical RNN with $h_t = f(x_t, h_{t-1})$, we also have $\rho_h^t = f(\rho_x^t, \rho_h^{t-1})$ where the updating function $f(\cdot)$ is parameterized by unitary matrices $U_x, U_h$ and real value $\lambda$. We posit that QRNN is potentially superior to RNN. A density matrix characterizes a probability measure on the Hilbert Space by defining a probability value to every pure state. This allows QRNN to better render the uncertainties in the conversational context with its hidden unit. Under this view, the hidden unit of a classical RNN can be seen as a pure state collapsing from the probability measure, with uncertainties removed. Moreover, the unitary transformation ensure zero information loss, since unitary transformation is an *entropy-preserving operation*, i.e.

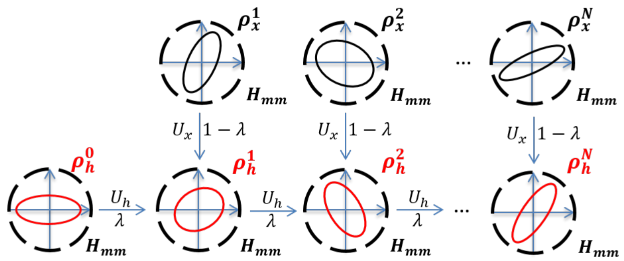


Figure 4: Diagram of the Quantum-like Recurrent Neural Network. With initial value $\rho_0^h$, the hidden density matrix $\rho_t^h$ for each time stamp $t$ is iteratively produced by Eq. 6.

$S(\rho) = S(U\rho U^*), \forall UU^* = I$[4]. This means QRNN has a strong potential in memorizing the context information. In comparison, there is always information loss in a classical RNN in the step of multiplying by the weight matrix.

Since the inputs and outputs are density matrices, QRNN could be stacked on top of one another to better capture the conversational context. The output of layer $l$ is generated by the QRNN with its previous layer as input, i.e. $\{\rho_l^t\} = QRNN(\{\rho_{l-1}^t\})$. In this work, however, we only use one layer of QRNN to construct the quantum-like contextual representation $\{\rho_c^t\}$. The exploitation of multi-layered QRNN is left for future work.

**Measurement and Collapse** After evolution, a sequence of $d$-dimensional states $\{\rho_t^c\}$ are obtained. A global observable $O$ is introduced to measure the emotional state of each utterance. The mutually orthogonal eigenstates form a d-dimensional unitary matrix $M$. After the measurement, a $d$-dimensional probability distribution is calculated, denoting the likelihood the state collapses onto the corresponding eigenstates. Since $d$ is often far greater than the number of emotions, the eigenstates could not be explicitly interpreted as emotional states. Instead, each emotion correspond to a high-dimensional subspace in the Hilbert Space $H_{mm}$ spanned by the eigenstates. In practice, we map the probability distribution to emotion label with a neural network with one hidden layer.

### Network Training

End-to-end training of a quantum-inspired complex-valued network has been discussed in (Li, Wang, and Melucci 2019). In this work, special care should be taken for the unitary matrices in the QRNN and measurement layers. In order to meet the unitary constraint throughout the training process, we employ the Riemannian approach proposed in (Wisdom et al. 2016) to update unitary matrices.The process of updating a unitary matrix $X$ is given by

[4]Please check (Nielsen and Chuang 2011) for the introduction of density matrix entropy.

| Dataset | # dialogues | | | # utterances | | |
|---|---|---|---|---|---|---|
| | train | dev | test | train | dev | test |
| IEMOCAP | 96 | 24 | 31 | 6808 | 1702 | 1623 |
| MELD | 1039 | 114 | 280 | 9989 | 1109 | 2610 |

Table 1: Distribution of training, test and validation sets for IEMOCAP and MELD.

$$G = \frac{\partial L}{\partial X} \tag{7}$$

$$A = G^H X - X^H G \tag{8}$$

$$\hat{X} = (I + \frac{lr}{2}A)^{-1}(I - \frac{lr}{2}A)X, \tag{9}$$

where $G$ is the general gradient, and the learning rate $lr$ controls to what extent $\hat{X}$ deviates from $X$. It can be proved that $(I + \frac{lr}{2}A)^{-1}(I - \frac{lr}{2}A)$ is always a unitary matrix, so the update value $\hat{X}$ is also a unitary matrix. However, the inverse of a complex matrix $(I + \frac{lr}{2}A)^{-1}$ is not directly tractable in a deep learning toolbox. To tackle this problem, we decompose the complex matrix inverse $Z = A + Bi$ into its real and imaginary parts:

$$(A+Bi)^{-1} = (A+BA^{-1}B)^{-1} - (B+AB^{-1}A)^{-1}i \tag{10}$$

In this way the inverse of a complex matrix can be implemented in a common deep learning toolbox. In this work, we implemented a separate optimizer for unitary parameters in Pytorch.

## Experiments

### Datasets

We evaluate our model on two benchmarking datasets, IEMOCAP (Busso et al. 2008b) and MELD (Poria et al. 2019b). IEMOCAP contains videos of dyadic conversations among 10 speakers under diverse scenarios. MELD is a multi-party conversation dataset crawled from the Friends TV series. For a fair comparison, we use the publicly available pre-trained utterance features provided by the authors of DialogueRNN (Majumder et al. 2019), available at github [5]. IEMOCAP has a 100-dim textual feature vector, a 512-dim visual feature vector and 100-dim acoustic feature vector for each utterance. MELD has 600-dim textual features and 300-dim acoustic features [6]. Table 1 shows the statistics of utterances and dialogues for both datasets. The emotion labels for IEMOCAP are *Happy, Sad, Neutral, Angry, Excited, Frustrated*. The emotion labels for MELD are *Fear, Sad, Neutral, Angry, Surprise, Disgust, Joy*.

### Models

We include a great variety of existing models in the experiment. For monologue models, Memory Fusion Network

[5]https://github.com/SenticNet/conv-emotion

[6]MELD pre-trained visual features are not publicly available.

| Model | IEMOCAP | | MELD | |
|---|---|---|---|---|
| | Acc | F1 | Acc | F1 |
| BC-LSTM | 58.66 | 58.45 | 59.80 | 57.60 |
| MFN | 58.22 | 58.11 | 60.65 | 57.79 |
| MulT | 59.52 | 59.47 | 60.54 | 57.94 |
| CMN | 59.46 | 59.33 | - | - |
| ICON | 56.93 | 56.89 | 59.96 | 57.18 |
| DialogueRNN | **62.42** | **60.48** | 57.73 | 55.58 |
| QMNN | 60.84 (-2.54) | 59.88 (-1.00) | **60.81** (0.26) | **58.00** (0.10) |

Table 2: Performances of conversational emotion recognition models on IEMOCAP and MELD in percentage(%). The best values among all models are in bold. The values in the parentheses are the relative differences in percentage(%) between QMNN variants and the best existing model. Acc = Accuracy.

| Model | Happy | | Sad | | Neutral | | Angry | | Excited | | Frustrated | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | F1 | P | F1 | P | F1 | P | F1 | P | F1 | P | F1 |
| BC-LSTM | 38.52 | 35.34 | 78.26 | 67.13 | 55.56 | 51.60 | 54.35 | 62.50 | 67.01 | 66.44 | 55.22 | 60.40 |
| MFN | 42.73 | 37.01 | 73.08 | 67.11 | 51.20 | 53.25 | 63.69 | **63.31** | 64.26 | 60.14 | 56.11 | 61.26 |
| MulT | 43.55 | **40.30** | 72.53 | 70.71 | 52.46 | 52.54 | **66.22** | 61.64 | **69.17** | 65.13 | 55.46 | 61.08 |
| CMN | 42.02 | 38.02 | 77.78 | 63.94 | 55.04 | **56.64** | 59.02 | 61.19 | 67.17 | 63.12 | **56.90** | 63.33 |
| ICON | 37.90 | 35.07 | 75.58 | 62.35 | 50.89 | 54.91 | 62.73 | 61.03 | 62.89 | 58.02 | 55.65 | 60.88 |
| DialogueRNN | **86.36** | 22.89 | **81.90** | **76.00** | **62.12** | 54.60 | 64.89 | 56.00 | 64.06 | **72.04** | 52.87 | **63.60** |
| QMNN | 41.35 | 39.71 | 72.86 | 68.30 | 54.11 | 55.29 | 65.38 | 62.58 | 66.04 | 66.71 | 55.56 | 62.19 |

Table 3: Performances of models on IEMOCAP in percentage(%). The best performance values among all models are in bold. P = Precision.

(MFN) (Zadeh et al. 2018b) and Multimodal Transformer (MulT) (Tsai et al. 2019b) are adapted to conversational context: the word-level inputs in the monologue setting are changed to input utterance features, and an output is yielded at each time stamp. For dialogue models, contextual LSTM model (BC-LSTM) (Poria et al. 2017) memory-based models including CMN (Hazarika et al. 2018b) and ICON (Hazarika et al. 2018a), and state-of-the-art model (DialogueRNN) (Majumder et al. 2019) are included. All these four models concatenate multimodal features at utterance level. We exclude DialogueGCN (Ghosal et al. 2019) from the experiment due to its instability under the multimodal setting.

The main evaluation metrics are the average accuracy and F1 scores over all emotions. In addition, the precision and F1 values for each emotion are calculated as a reference. For a fair comparison, a grid search for the best hyper-parameters is conducted for all models. At each search the model is trained for 100 epochs and the model with the lowest validation loss is chosen. The best-performed model on the test set out of 50 searches is taken as the model performance.

On both datasets, QMNN hyper-parameters are searched within embedding dimensions $d \in \{100, 120, 140, 160, 180, 200\}$, the size of last hidden layer in $\{32, 48, 64, 80\}$. Stochastic gradient descent (SGD) is used as the optimizer with a learning rate $lr \in \{0.001, 0.002, 0.005, 0.008\}$. The unitary matrix training algorithm is also modified to an SGD fashion, where the general gradient $G$ in Eq. 7 is replaced by the SGD gradient. The learning rate *unitary-lr* for updating the unitary matrix varies in $\{0.001, 0.002, 0.005, 0.008\}$. The batch size $bs$ varies in $\{24, 48, 96\}$ for MELD and $\{4, 8, 16\}$ for IEMOCAP in proportion to the dataset scale. The dropout rate for the last hidden layer varies in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. We set the number of parties $K = 1$ for MELD and $K = 2$ for IEMOCAP. Since the total number of speakers (actors) is huge (250 in training, 46 in validation, 48 in test), speaker-dependent encoding may suffer from data sparsity issue for MELD. For the same reason, CMN is removed from the MELD experiment. IEMOCAP has a male speaker and a female speaker in each conversation, so learning two set of frequencies and initial phases may capture the gender-related factors that influence the emotion. The experiments are run on a Linux server with one NVidia Tesla V100 Graphic card. The codes are implementation in PyTorch, publicly available on GitHub [7].

# Results and Discussion

## Overall Performance

The overall effectiveness results on IEMOCAP and MELD are shown in Table 2, while the per-emotion values are presented in Table 3 and 4 respectively[8]. The proposed QMNN attains the best overall F1 and Accuracy scores on MELD by a tiny margin, and beats all remaining models but slightly

[7] github.com/qiuchili/diasenti.git

[8] All models have zero values on *Fear* and *Disgust*.

| Model | Sad | | Neutral | | Angry | | Surprise | | Joy | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | F1 | P | F1 | P | F1 | P | F1 | P | F1 |
| BC-LSTM | 32.08 | 21.66 | **73.13** | 76.20 | 42.36 | 44.01 | 45.09 | 49.76 | 51.07 | 52.13 |
| MFN | 27.86 | 22.41 | 70.74 | 76.77 | **47.40** | 43.22 | **58.55** | 47.68 | 48.19 | **53.33** |
| MulT | 36.17 | **22.52** | 71.74 | 76.46 | 42.75 | **46.64** | 49.61 | 47.50 | 51.82 | 52.40 |
| ICON | **43.10** | 18.80 | 73.03 | 76.54 | 44.63 | 41.37 | 41.34 | **50.64** | 48.57 | 51.46 |
| DialogueRNN | 24.00 | 15.58 | 72.13 | 74.70 | 34.51 | 46.09 | 41.13 | 47.76 | 51.62 | 49.48 |
| QMNN | 24.30 | 16.50 | 71.23 | **77.00** | 42.86 | 43.17 | 45.81 | 49.76 | **53.48** | 52.08 |

Table 4: Performances of models on MELD in percentage(%). The best performance values among all models are in bold. P = Precision.

underperforms DialogueRNN on IEMOCAP. We acknowledge that no significance differences between our model and existing models are observed, and the results therefore suggest comparable effectiveness to existing models.

## Ablation Study

To investigate the effect of the introduced quantum components, an ablation study is carried out. To examine the quantum mixture component, we build *QMNN-concat*, which computes utterance amplitudes by projecting the concatenation of unimodal features, and *QMNN-realmix* that ignores the phases of an utterance, with a doubled projection dimension. QRNN is contrasted by a classical GRU, and no recurrent structure, termed as *QRNN-crec* and *QRNN-norec* respectively. Finally, we consider two variants of our quantum measurement, a semantic measurement (Li, Wang, and Melucci 2019) (a.k.a *QMNN-seamea*) with the same number of eigenstates as QMNN and a fully connected network with one hidden layer on flattened density matrix (a.k.a *QMNN-flatten*). *QMNN-seamea* has a pre-defined number of eigenstates that are not necessarily orthogonal to each other, and hence trained by classical backpropagation algorithm.

| Model | Acc | F1 |
|---|---|---|
| QMNN | 0.6080 | 0.5800 |
| **Preparation** | | |
| QMNN-concat | 0.5904 (-2.89%) | 0.5623 (-3.05%) |
| QMNN-realmix | 0.5938 (-2.33%) | 0.5716 (-1.45%) |
| **Evolution** | | |
| QMNN-norec | 0.5945(-2.22%) | 0.5684 (-2.00%) |
| QMNN-crec | 0.5889 (-3.14%) | 0.5686 (-1.97%)) |
| **Measurement** | | |
| QMNN-seamea | 0.5938 (-2.34%) | 0.5673 (-2.19%) |
| QMNN-flatten | 0.5959 (-1.99%) | 0.5700 (-1.72%) |

Table 5: Ablation Study on MELD. Values in parentheses are the relative differences from QMNN.

Table 5 exhibits a performance drop after each quantum component is replaced by its classical counterparts. The quantum mixture step effectively fuses multimodal data and integrates the utterance order, which agrees with the argument in (Wang et al. 2020) on the compatibility of order information with the phase-amplitude assignment mechanism. Furthermore, additional conversational contextual information is captured by QRNN as it yields a performance gain over *QMNN-norec*. The increase over classical RNN suggests the superiority in capturing ambiguities with the density matrix hidden unit of QRNN. Finally, compared to previous works (Li, Wang, and Melucci 2019; Wang et al. 2019), QMNN proposed an authentic quantum measurement with mutually orthogonal operators, leading to no performance drop.

## Efficiency Analysis

To satisfy the numerical constraints for the quantum components, we introduced a special training algorithm for complex-valued unitary matrix. The efficiency bottleneck falls on the matrix inverse (Eq. 10), which is of the same computational complexity degree as matrix multiplication $(O(n^3) - O(n^{2.373}))$[9]. The unitary matrix training is expected to moderately increase the training time. To examine this argument, we compare the average training time per batch of two QMNN variants with no recurrent structures, one with quantum measurement and the other with semantic measurement. Out of 50 size-16 batches, an average of 0.075s time difference per batch is observed, suggesting an affordable efficiency of the unitary matrix training.

# Conclusions

This work provides a novel quantum view to the conversational emotion recognition problem. A holistic quantum-inspired network is constructed to fuse multimodal data and build conversational context, and identify per-utterance emotion on its basis. The design of the network and the adoption of unitary matrix training ensures the authenticity of quantum analogy, which is not at a tremendous sacrifice of effectiveness or efficiency as illustrated in a comprehensive comparison with state-of-the-art models on two benchmarking datasets.

# Acknowledgements

This study is supported by the Quantum Information Access and Retrieval Theory (QUARTZ) project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 721321.

[9]Computational complexity of mathematical operations, Wikipedia.

## References


Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246. Melbourne, Australia: Association for Computational Linguistics.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv:1705.09406 [cs]* URL http://arxiv.org/abs/1705.09406. ArXiv: 1705.09406.

Barezi, E. J.; and Fung, P. 2019. Modality-based Factorization for Multimodal Fusion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 260–269. URL https://www.aclweb.org/anthology/W19-4331/.

Born, M. 1926. Zur Quantenmechanik der Sto\s svorgänge. *Zeitschrift für Physik* 37(12): 863–867. ISSN 0044-3328. doi:10.1007/BF01397477.

Busemeyer, J. R.; and Bruza, P. D. 2012. *Quantum Models of Cognition and Decision*. New York, NY, USA: Cambridge University Press, 1st edition. ISBN 978-1-107-01199-1.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008a. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42(4): 335–359. ISSN 1574-020X, 1574-0218. doi:10.1007/s10579-008-9076-6. URL http://link.springer.com/10.1007/s10579-008-9076-6.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008b. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4): 335.

Eyben, F.; Wöllmer, M.; Graves, A.; Schuller, B.; Douglas-Cowie, E.; and Cowie, R. 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *J Multimodal User Interfaces* 3(1): 7–19. ISSN 1783-8738. doi:10.1007/s12193-009-0032-6. URL https://doi.org/10.1007/s12193-009-0032-6.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 154–164. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1015. URL https://www.aclweb.org/anthology/D19-1015.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2594–2604. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1280. URL https://www.aclweb.org/anthology/D18-1280.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018b. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1193.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1): 221–231. ISSN 1939-3539. doi:10.1109/TPAMI.2012.59.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics. doi:10.3115/v1/D14-1181. URL https://www.aclweb.org/anthology/D14-1181.

Li, Q.; Wang, B.; and Melucci, M. 2019. CNM: An Interpretable Complex-valued Network for Matching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4139–4148. Minneapolis, Minnesota: Association for Computational Linguistics.

Liang, P. P.; Liu, Z.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 150–161. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1014. URL https://www.aclweb.org/anthology/D18-1014.

Liang, P. P.; Liu, Z.; Tsai, Y.-H. H.; Zhao, Q.; Salakhutdinov, R.; and Morency, L.-P. 2019. Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1569–1576. Florence, Italy: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1152.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256. Melbourne, Australia: Association for Computational Linguistics.

Mai, S.; Hu, H.; and Xing, S. 2019. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 481–492. Florence, Italy: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1046.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *arXiv:1811.00405 [cs]* URL http://arxiv.org/abs/1811.00405. ArXiv: 1811.00405.

Nielsen, M. A.; and Chuang, I. L. 2011. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. New York, NY, USA: Cambridge University Press, 10th edition. ISBN 978-1-107-00217-3.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. *1* 33: 6892–6899. ISSN 2374-3468. doi:10.1609/aaai.v33i01.33016892. URL https://aaai.org/ojs/index.php/AAAI/article/view/4666.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–883. Vancouver, Canada: Association for Computational Linguistics. doi:10.18653/v1/P17-1081. URL https://www.aclweb.org/anthology/P17-1081.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019a. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019b. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1050.

Ringbauer, M. 2017. *Exploring Quantum Foundations with Single Photons*. Springer Theses. Springer International Publishing. ISBN 978-3-319-64987-0. doi:10.1007/978-3-319-64988-7. URL https://www.springer.com/gp/book/9783319649870.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019a. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. Florence, Italy: Association for Computational Linguistics.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019b. Multimodal Transformer for Unaligned Multimodal Language Sequences. *arXiv preprint arXiv:1906.00295* .

Wang, B.; Li, Q.; Melucci, M.; and Song, D. 2019. Semantic Hilbert Space for Text Representation Learning. In *The World Wide Web Conference*, WWW '19, 3293–3299. New York, NY, USA: ACM. ISBN 978-1-4503-6674-8. doi:10.1145/3308558.3313516. Event-place: San Francisco, CA, USA.

Wang, B.; Zhao, D.; Lioma, C.; Li, Q.; Zhang, P.; and Simonsen, J. G. 2020. Encoding word order in complex embeddings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL https://openreview.net/forum?id=Hke-WTVtwr.

Wisdom, S.; Powers, T.; Hershey, J.; Le Roux, J.; and Atlas, L. 2016. Full-Capacity Unitary Recurrent Neural Networks. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*, 4880–4888. Curran Associates, Inc. URL http://papers.nips.cc/paper/6327-full-capacity-unitary-recurrent-neural-networks.pdf.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1103–1114. Association for Computational Linguistics. doi:10.18653/v1/D17-1115.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory Fusion Network for Multi-view Sequential Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 5634–5641.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018b. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018c. Multi-attention Recurrent Network for Human Communication Comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 5643–5649.

Zhang, Y.; Song, D.; Li, X.; Zhang, P.; Wang, P.; Rong, L.; Yu, G.; and Wang, B. 2020. A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion* 62: 14–31. ISSN 1566-2535. doi:10.1016/j.inffus.2020.04.003. URL http://www.sciencedirect.com/science/article/pii/S1566253520302554.