Information Fusion xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis

Q1 Yazhou Zhang [a], Dawei Song [b,c,*], Xiang Li [d], Peng Zhang [e], Panpan Wang [e], Lu Rong [a], Guangliang Yu [f], Bo Wang [e]

[a] Softwware Engineering College, Zhengzhou University of Light Industry, 450002, No.136 Science Avenue, Zhengzhou, Henan Province, P.R.China
[b] School of Computer Science and Technology, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing 100081, P.R. China
[c] School of Computing and Communications, the Open University, Walton Hall, Milton Keynes, MK7 6AA. United Kingdom
[d] Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), 250000, No.19 Keyuan Road, Lixia District, Jinan, P.R. China
[e] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, 300350, No.135 Yaguan Road, Jinnan District, Tianjin, P.R.China
[f] Meituan-Dianping Group, Hengdian Building, No. 4 Wangjing East Road, Chaoyang District, Beijing, P.R. China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Interactive dynamics
Human conversation
Quantum theory
Long short-term memory (LSTM) network

ABSTRACT

Sentiment analysis in conversations is an emerging yet challenging artificial intelligence (AI) task. It aims to discover the affective states and emotional changes of speakers involved in a conversation on the basis of their opinions, which are carried by different modalities of information (e.g., a video associated with a transcript). There exists a wealth of intra- and inter-utterance interaction information that affects the emotions of speakers in a complex and dynamic way. How to accurately and comprehensively model complicated interactions is the key problem of the field. To fill this gap, in this paper, we propose a novel and comprehensive framework for multimodal sentiment analysis in conversations, called a quantum-like multimodal network (QMN), which leverages the mathematical formalism of quantum theory (QT) and a long short-term memory (LSTM) network. Specifically, the QMN framework consists of a multimodal decision fusion approach inspired by quantum interference theory to capture the interactions within each utterance (i.e., the correlations between different modalities) and a strong-weak influence model inspired by quantum measurement theory to model the interactions between adjacent utterances (i.e., how one speaker influences another). Extensive experiments are conducted on two widely used conversational sentiment datasets: the MELD and IEMOCAP datasets. The experimental results show that our approach significantly outperforms a wide range of baselines and state-of-the-art models.

1. Introduction

Multimodal sentiment analysis has been a core research topic in artificial intelligence (AI)-related areas, e.g., affective computing, information fusion, and multimodal interaction [1–6]. Unlike traditional text-based analysis, multimodal sentiment analysis requires both the application of multimodality representation techniques and information fusion techniques [7–10], such as feature-level [11,12], decision-level [13] and hybrid fusion [14] techniques. Most existing multimodal sentiment analysis approaches focus on identifying the polarity of people's opinions, which are posted in social media platforms, e.g., YouTube [15], Flickr [13], Getty Images [16], and MOSI [12]. The multimodal documents used in these studies are usually in the form of individual narratives, without involving interactions among speakers or writers.

The recent advancement of internet and instant messaging services, such as Skype, Line and WeChat, has produced a massive volume of multimodal records of communications between humans. Such data are a rich source of information, including that of sentiments or opinions, which often evolve during conversations [17,18]. This advancement brings forth a new challenge of judging the evolving sentiment polarities of different people in a conversational discourse. Therefore, research on conversational sentiment analysis has attracted increasing attention from both academia and industry [19–21].

Multimodal sentiment analysis in conversations (also called conversational multimodal sentiment analysis) aims to detect the affective states of multiple speakers and study the sentimental change of each speaker in the course of the interaction. Different from the previous multimodal sentiment analysis approaches, which focus on describing the interactions between different modalities, the interaction dynamics in

* Corresponding author.
E-mail address: dawei.song2010@gmail.com (D. Song).

https://doi.org/10.1016/j.inffus.2020.04.003
Received 13 June 2019; Received in revised form 2 April 2020; Accepted 11 April 2020
Available online xxx
1566-2535/

Please cite this article as: Y. Zhang, D. Song and X. Li et al., A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis, Information Fusion, https://doi.org/10.1016/j.inffus.2020.04.003

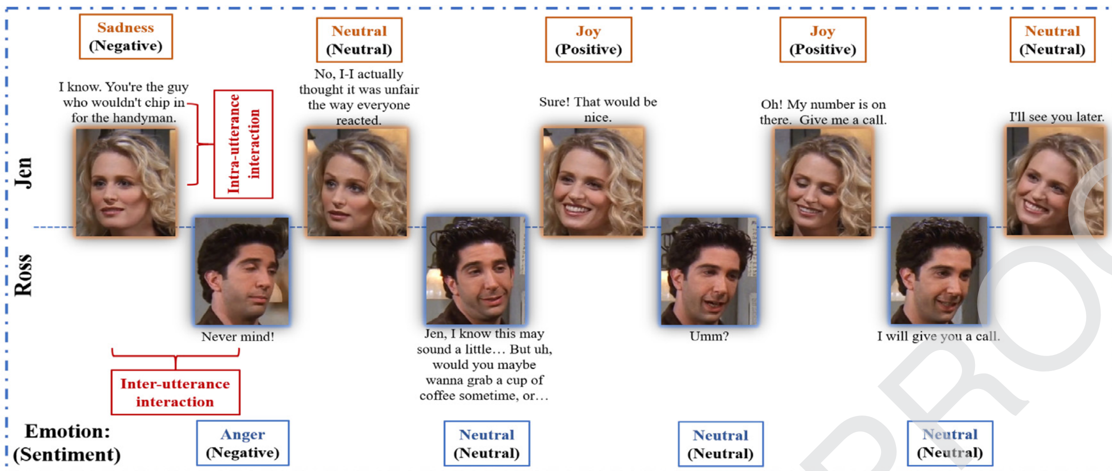

Fig. 1. Two interaction dynamics in a conversation. Red and blue are used to show the emotional shifts of Jen and Ross, respectively.

conversations are more complex, involving intra- and inter-utterance interactions. Intra-utterance interaction refers to the correlation between different modalities within one utterance, such as the mutual influence, joint representation, and decision fusion. Inter-utterance interaction involves repeated interactions among speakers, resulting in the exchange of ideas and having an effect on one another. Fig. 1 provides an example from the MELD dataset [22] that shows the presence of these two patterns in a conversation. From Fig. 1, we can notice that *Jen* and *Ross*'s affective states change dynamically because of intra- and inter-utterance interactions.

There has been a growing body of literature on conversational sentiment analysis. For instance, Welch et al. [19] proposed a neural model using longitudinal dialogue data for two dialogue prediction tasks: next message prediction and response time prediction. However, their work did not involve sentiment analysis. Ojamaa et al. [23] developed a lexicon-based technology to extract the speaker's attitude from conversational texts. However, they neglected the interaction information and used a text dataset rather than a multimodal dataset. Bhaskar et al. [24] proposed combining acoustic and textual features for emotion classification of audio conversations. Although they enhanced the efficiency of emotion classification, they did not consider interactions among speakers, i.e., inter-utterance interactions. Hazarika et al. [21] proposed a conversational memory network that uses contextual information from the conversation history to recognize emotions in dyadic dialogue videos. However, as they admitted, the work was limited to dyadic conversation scenarios and might not be applicable to multiparty conversations [22]. These previous methods treated utterances as independent and ignored the order of the utterances. Poria et al. [20] proposed a contextual h-LSTM network that takes the sequence of utterances in a video as input and extracts contextual features by modeling the dependencies among the input utterances. They also created a multimodal multiparty conversational dataset, namely, the Multimodal EmotionLines Dataset (MELD), to facilitate the development of conversational sentiment analysis.

In recent years, quantum theory (QT), as a mathematical formalism to model the complex interactions and dynamics in quantum physics, has been adopted for constructing text representations in various information retrieval (IR) and NLP tasks [13,25–27]. For instance, the quantum language model (QLM) [25] represents a query or document as a density matrix on a quantum probability space, which could evolve with respect to the user search/dialogue session through matrix transformations [28]. Based on the QLM, density matrix-based metrics can be computed to serve as ranking functions. Neural network-based QLM (NNQLM) [29] builds an end-to-end network for question answering (QA) to jointly model a question-answer pair based on their density matrix representations. Motivated by this work, a quantum-like interactive network model was proposed to recognize the sentiment polarity of each conversation [30]. Such QT-based models could be considered a generalization of traditional approaches in that they are capable of capturing inherent intricacies within interactions. These studies motivate us to explore the use of quantum theory as a theoretical basis for capturing the intra- and inter-utterance interaction dynamics, both of which are complex in nature.

In this paper, drawing upon the quantum theory formalism and the LSTM architecture, we propose a novel and comprehensive quantum-like multimodal network (QMN) framework, which jointly models the intra- and inter-utterance interaction dynamics by capturing the correlations between different modalities and inferring dynamic influences among speakers. Fig. 3 illustrates the QMN framework. First, the QMN extracts and represents multimodal features (e.g., text and images) for all utterances in one video using a density matrix-based CNN (DM-CNN) subnetwork and takes them as inputs. Second, inspired by quantum measurement theory, the QMN introduces a strong-weak influence model to measure the influences among speakers across utterances and feeds the resulting influence matrices into the QMN by incorporating them into the output gate of each LSTM unit. Third, with textual and visual features as inputs, the QMN employs two individual LSTM networks to obtain their hidden states, which are fed to the softmax functions to obtain the local sentiment analysis results. Finally, a multimodal decision fusion approach inspired by quantum interference is designed to derive the final decision based on the local results.

We have designed and carried out extensive experiments on two widely used conversational sentiment datasets (the MELD and IEMOCAP datasets) to demonstrate the effectiveness of the proposed QMN framework in comparison with a wide range of baselines, including two unimodal approaches, a feature-level fusion approach and a decision-level fusion approach, and five state-of-the-art multimodal sentiment analysis models. The results show that the QMN significantly outperforms all these comparative models.

The major innovations of the work presented in this paper are summarized as follows.

- We propose a quantum-like multimodal network framework, which leverages quantum probability theory within the LSTM architecture, to model both intra- and inter-utterance interaction dynamics for multimodal sentiment analysis in conversations.
- We propose a quantum interference-inspired multimodal decision fusion method to model the decision correlations between different modalities.
- We propose a quantum measurement-inspired strong-weak influence model to make better inferences about social influence among speakers than with previous methods.

The rest of this paper is organized as follows. Section 2 presents a brief review of the related work. Section 3 introduces the preliminaries of quantum probability theory. In Section 4, we describe the proposed quantum-like multimodal network framework in detail. In Section 5, we report the empirical experiments and analyze the results. Section 6 concludes the paper and points out future research directions.

2. Related work

Now, we present a brief review of the related work, including multimodal sentiment analysis and conversational sentiment analysis.

2.1. Multimodal sentiment analysis

Generally, multimodal sentiment analysis refers to the use of natural language processing, information fusion techniques, statistics or machine/deep learning methods to identify the subjective attitude of an author expressed in multimodal documents that may involve visual, audio and textual information [31,32]. An early example was Yoshitomi's integration approach to recognizing human emotions carried in voices and facial expressions [33]. Then, sentiment analysis began to be performed in a multimodal framework [34]. Similarly, Mehrabian [35] argued that when judging people's affective states, one mainly relies on facial expressions and vocal intonations.

Building on these works, Sebe et al. [36] performed emotion recognition by combining cues from facial expressions and vocal information. Morency [37] addressed for the first time the task of trimodal sentiment analysis and showed that it could benefit from the joint exploitation of visual, audio and textual modalities. Mihalcea et al. [38] created a multimodal dataset consisting of sentiment-annotated utterances extracted from video reviews. Zhang et al. [13] explored the use of quantum theory (QT) to model a sentiment analysis task and proposed a quantum-inspired multimodal sentiment analysis (QMSA) model. However, they were unable to deal with the interactions between different contextual utterances. Inspired by them, Gkoumas and Song [39] exploited quantum-like interference in decision fusion for ranking multimodal documents. Li [40] tried to fuse multimodal data with complex-valued neural networks, motivated by the theoretical link between neural networks and quantum theory. They [41] also introduced a work in progress that targeted building a multimodal representation under quantum inspiration. However, they only focused on the interactions between different modalities. Moreover, there have been many emerging studies on other NLP tasks, such as information retrieval [13] and text classification [42].

Currently, a large body of research on multimodal sentiment analysis is performed from a multimodal learning perspective. There are an increasing number of studies that have used deep neural networks [6,43,44]. For instance, You et al. [45] proposed a progressively trained convolutional neural network (CNN) for visual sentiment analysis and achieved state-of-the-art performance. Furthermore, they proposed a cross-modality consistent regression (CCR) model to analyze Getty Images and Twitter multimedia content [16]. Zadeh et al. [11] introduced a tensor fusion network to fuse audio and visual features. Chen et al. [46] proposed a gated multimodal embedding LSTM with temporal attention model to alleviate the difficulties of fusion. Poria et al. [47] introduced an attention-based network for improving both context learning and dynamic feature fusion. Huang et al. [48] proposed a deep multimodal attentive fusion approach to exploit discriminative features and the internal correlation between visual and semantic contents. Kumar et al. [1] proposed a multimodal framework that can fuse EEG signals, product descriptions and brand reviews to predict ratings given by consumers. Poria et al. [12] published an overview of multimodal sentiment analysis and developed three deep learning-based architectures as baselines. Their team also considered the correlations between sarcasm detection and sentiment analysis in multitask learning [49]. Yu and Jiang [50] proposed a multimodal BERT model to obtain target-sensitive textual and visual representations for the task of target-oriented multimodal sentiment classification. Verma et al. [51] first proposed a deep network to extract the common information from the multimodal representations and thus designed another model to mine the modality-specific information for multimodal sentiment analysis. Xu et al. [52] proposed a new subtask, named aspect-based multimodal sentiment analysis, which could be seen as the combination of aspect-level sentiment analysis and multimodal sentiment analysis. They also designed a multi-interactive memory network model for this subtask. Considering the problem of "missing modality", Fortin et al. [53] proposed a multimodal model that leveraged a multitask framework to enable the use of training data composed of an arbitrary number of modalities, and it could also perform predictions with missing modalities. Chaturvedi et al. [54] employed deep learning-based models to extract features from each modality and then mapped them into a common sentiment space that had been clustered into different emotions via a convolutional fuzzy sentiment classifier. Huddar and Sannakki [55] summarized the latest computational approaches used in multimodal sentiment analysis and the associated challenges. Dumpala et al. [56] considered the special scenario where both modalities were available during training but only one modality was available during testing and combined deep canonical correlation analysis with cross-modal autoencoders.

2.2. Conversational sentiment analysis

Traditional sentiment analysis research mainly focuses on identifying the polarities of personal reviews. With the increasing popularity of social networks, conversational sentiment analysis has attracted an increasing attention.

Elise et al. [57] presented an approach for the detection of both the topic and sentiment of a user's utterances from transcribed speech. They obtained the sentiment scores based on sentiment rules. Yang et al. [17] proposed a segment-level joint topic-sentiment model (STSM) to estimate fine-grained sentiments for online review analysis. Mahata et al. [58] trained a shallow convolutional neural network (CNN) model based on annotated Twitter responses for detecting personal exposure. Contrary to our model, they ignored interactions between authors. Maghilnan et al. [59] performed a sentiment analysis on speaker-discriminated speech transcripts to detect the emotions of the individual speakers involved in a conversation using machine learning classifiers. Realizing the difficulty of gaining insights from long conversations, Hoque and Carenini [60] developed a visual exploratory text analytic system that integrates interactive visualization with text mining techniques. Mazzocut et al. [61] manually analyzed people's opinions, which were collected from web conversations. Due to the limited availability of sentiment-annotated interactive text datasets, Bothe et al. [62] had to use the VADER sentiment analysis tool [60] to autoannotate the sentiment labels of two spoken interaction corpora for training. Motivated by the above studies, Huijzer et al. [63] performed an affective analysis of emails and collected an email sentiment dataset. They noticed, but did not model, the interaction between the customer support agent and a customer. From a sociological perspective, Aznar and Tenenbaum [64] employed a meta-analysis to compare gender differences in the frequency of mother-child emotion talk and the moderators of these differences.

Unlike the aforementioned studies, Hazarika et al. [21] proposed a conversational memory network, which leveraged contextual information from the conversation history, to recognize utterance-level emotions. However, their work was limited to dyadic conversation understanding. Majumder el al[65]. described a DialogueRNN model that kept track of the individual party states throughout the conversation and used this information for emotion classification in conversations. Poria et al. [20] proposed an LSTM-based model that was able to capture contextual information of utterances from their surroundings in a video, thus aiding the classification process. Moreover, Poria et al. [22] created the first multimodal multiparty conversational dataset, namely, the Multimodal EmotionLines Dataset (MELD), to facilitate the development of conversational sentiment analysis. Zhang et al. [66] treated each utterance and each speaker in each conversation as a node and designed a conversational graph-based convolutional neural network to model contextual dependency. Zhong et al. [67] also attempted to address this problem and proposed a knowledge-enriched transformer (KET) that used a context-aware affective graph attention mechanism to learn external contextual knowledge. Zhang et al. [30] designed a quantum-inspired interactive network (QIN) model for textual conversational sentiment analysis and showed its effectiveness on the MELD and IEMOCAP datasets. However, they did not take the interactions among different modalities into consideration. Rebiai et al. [68] presented one submission at SemEval-2019 Task 3: EmoContext. The task consisted of classifying a textual dialogue into one of four emotion classes: happy, sad, angry or other. They provided a series of strong baseline approaches for supporting the development of sentiment analysis of conversations.

In summary, the two aforementioned types of studies have made good progress in multimodal sentiment analysis and motivated our work. The existing research is mainly focused on leveraging intra-utterance interactions, e.g., learning relations between words and extracting effective features, to help judge sentiment. A few studies in the last two years have attempted to implicitly train models to learn the interactions between utterances using deep neural networks. However, to the best of our knowledge, they have not yet systematically taken into account the three kinds of interactions (i.e., interactions between terms, interactions among speakers and interactions between modalities) in a unified framework, as we aim to address in this paper.

In this paper, we aim to take a fresh look at the nature of complex interactions from the perspective of quantum theory and establish an integrated theoretical system of quantum-like interaction modeling. As major parts of the theoretical system of quantum-like interaction modeling, the QMSA model [13] and the QIN model [30] are merged together under the same subject. Finally, under the guidance of the theoretical system, we propose a principled, theoretical framework to model both intra- and inter-utterance interactions that are complex and dynamic. The framework will draw upon the formalisms of quantum probability theory, which is a generalization of classical probability theory and is designed to describe the behaviors of microscopic particles in quantum physics, which are also dynamic and complex in nature.

3. Quantum theory preliminaries

Quantum probability theory [69] aims at interpreting the mathematical foundations of quantum theory, which is based on linear algebra. This section gives a brief introduction to some basic concepts, quantum measurement and quantum interference formalisms.

3.1. Basic notations and concepts

Quantum probability theory [69] aims at interpreting the mathematical foundations of quantum theory, which is based on linear algebra. This section gives a brief introduction to some basic concepts, quantum measurement and quantum interference formalisms.

3.2. Basic notations and concepts

In quantum theory, quantum probability space is naturally encapsulated in an infinite Hilbert space [70] (which is a complete vector space possessing the structure of an inner product), denoted by $\mathbb{H}$. In line with previous quantum-inspired models [26,27,29], we restrict our problem to vector spaces over real numbers in $\mathbb{R}$ and leave the possible extension to complex numbers as one direction of future work.

With Dirac's notation, a state vector or a wave function, φ, can be expressed as a ket $|\varphi\rangle$, and its transpose can be expressed as a bra $\langle\varphi|$. In Hilbert space, any n-dimensional vector can be represented in terms of a set of basis vectors, $|\varphi\rangle = \sum_{i=1}^{n} a_i |e_i\rangle$, as can the wave function. Given two state vectors $|\varphi_1\rangle$ and $|\varphi_2\rangle$, the inner product between them is denoted by $\langle\varphi_1|\varphi_2\rangle$. Similarly, the Hilbert space representation of the wave function is recovered from the inner product $\varphi(x) = \langle x|\varphi\rangle$.

In quantum probability theory, an event is defined as a subspace of Hilbert space, which is represented by any orthogonal projector Π. Assuming $|u\rangle$ is a unit vector; i.e., $\|\vec{u}\|_2 = 1$, the projector Π in the direction u is written as $|u\rangle\langle u|$. $\rho = \sum_i p_i |u\rangle\langle u|$ represents a density matrix. The density matrix ρ is symmetric, positive semidefinite, $\rho = \rho^T$, where $\rho \geq 0$, and has a trace of 1. The quantum probability measure μ is associated with the density matrix. It satisfies two conditions: (1) for each projector $|u\rangle\langle u|$, $\mu(|u\rangle\langle u|) \in [0, 1]$, and (2) for any orthonormal basis $\{|e_i\rangle\}$, $\sum_{i=1}^{n} \mu(|e_i\rangle\langle e_i|) = 1$. Gleason's theorem [71] has proven the existence of a mapping function $\mu(|u\rangle\langle u|) = tr(\rho|u\rangle\langle u|)$ for any vector $|u\rangle$.

In quantum theory, all the information contained in one system (which, in this paper, refers to each utterance) is represented by the probability distribution of the measurement results. These probabilities are obtained using a finite sequence of measurements on the system and are used to construct the state space[72]. Since the density matrix is equivalent to the state space, it describes all the information and properties of the system (utterance).

3.3. Quantum measurement

There are two types of quantum measurements (QMs), including ordinary (i.e., strong) and weak measurements. Quantum measurement describes the interactions between a quantum system and the measurement system. Strong measurement leads to the collapse of the quantum state, while weak measurement disturbs the quantum state very little. In QT, a quantum measurement process consists of two steps: (i) the quantum measurement device is weakly coupled to the quantum system being measured; (ii) the measurement device is strongly measured, and its collapsed state is referred to as the outcome of the measurement process.

Let $|\phi_d\rangle$ denote the wave function of the measurement device and represent the position basis. It can be written as:

$$|\phi_d\rangle = \int_x \phi(x)|x\rangle dx \tag{1}$$

$$\phi(x) = (2\pi\sigma^2)^{-\frac{1}{4}} e^{-x^2/4\sigma^2} \tag{2}$$

where x is the position variable of the measuring pointer. The initial state of the pointer variable is modeled by a Gaussian distribution centered at zero with variance σ^2 (denoted by Δ).

As an example, let S denote the quantum system being measured. Suppose $\hat{O}$ is observable in the system S. Taking $\hat{O} = \frac{\hbar}{2}|0\rangle - \frac{\hbar}{2}|1\rangle$, $\hbar$ is Planck's constant, which is the quantum of action. A quantum state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, in which α and β are the probability amplitudes, satisfies $|\alpha|^2 + |\beta|^2 = 1$. $|0\rangle$ and $|1\rangle$ are the eigenstates, and 0 and 1 are the eigenvalues of the two eigenstates. Then, the system and the measurement device can be entangled, which is formalized as:

$$\int_x \left[e^{-\frac{(x-0)^2}{4\sigma^2}} \alpha|0\rangle \otimes |x\rangle + e^{-\frac{(x-1)^2}{4\sigma^2}} \beta|1\rangle \otimes |x\rangle \right] dx \tag{3}$$

This function can be seen as a bimodal distribution with two modes (i.e., 0 and 1). More details on the entanglement process are provided in [73]. Next, we strongly measure the pointer of the measuring device. Supposing the pointer collapses to the vector $|x_0\rangle$, the system becomes new in the state:

$$\left[e^{-\frac{(x_0-0)^2}{4\sigma^2}} \alpha|0\rangle + e^{-\frac{(x_0-1)^2}{4\sigma^2}} \beta|1\rangle \right] \otimes |x_0\rangle \tag{4}$$

The eigenvalue x_0 could be anywhere around 0 or 1, or even further away. A smaller variance Δ indicates that the curve of the bimodal distribution will be taller and narrower. The value of x is tightly clustered around the two modes 0 and 1 (i.e., the two eigenvalues of the system),

Table 1
The parameter analysis for Equation 8.

Variance	Strong Measurement $\sigma < eigenvalue$		Weak Measurement $\sigma \geqslant eigenvalue$	
Position in Eq. 4	left side	right side	left side	right side
Supposing x_0 is approximately 1	$\frac{-(x-0)^2}{4\sigma^2} \to -\infty$ $e^{\frac{-(x-0)^2}{4\sigma^2}} \to 0$	$\frac{-(x-1)^2}{4\sigma^2} \to 0$ $e^{\frac{-(x-1)^2}{4\sigma^2}} \to 1$	$\frac{-(x-0)^2}{4\sigma^2} \to 0$ $e^{\frac{-(x-0)^2}{4\sigma^2}} \to 1$	$\frac{-(x-1)^2}{4\sigma^2} \to 0$ $e^{\frac{-(x-1)^2}{4\sigma^2}} \to 1$
Effect on the quantum state	collapsed to $\|1\rangle$		slightly biased	

which means that the probability of the system state collapsing to one of the eigenstates is very high. This type of measurement is called a strong measurement. A very large variance Δ indicates that the curve of the bimodal distribution will be flat and broad. The value of x is spread out and has a large uncertainty. The outcome of this measurement is the average over the probabilities of the two eigenvalues 0 and 1. Such measurement is called weak measurement. Hence, the higher the variance is, the weaker the measurement process.

Whether the quantum measurement (QM) is strong or weak is determined by Δ. If the pointer collapses to a value x_0 of approximately 1, it means that the amplitude to postselect $|0\rangle$ will be higher than the amplitude to postselect $|1\rangle$, and vice versa. Thus, the collapse of the pointer biases the system's vector. However, if σ is very large with respect to the eigenvalue of $\hat{O}$, the bias will be very small, and the outcome system's vector will be very similar to the original vector. A detailed analysis is shown in Table 1.

Strong measurement leads to the collapse of the quantum system, while weak measurement causes the quantum system to be slightly bias. QM provides a principled and effective mechanism to capture the inter-utterance interactions, which will be detailed in Section 4.3.1.

3.4. Preliminaries of quantum interference

The double-slit interference experiment [74], as shown in Fig. 2, is a demonstration that a single photon initially emitted as a particle goes through two slits simultaneously and interferes with itself as a wave. In QT, the wave function $\varphi(x)$ is a probability amplitude function of position x, which is used to interpret this experiment. The state of the photon is a superposition of the state of slit 1 and slit 2, which can be formulated as

$$\varphi_p(x) = \alpha\varphi_1(x) + \beta\varphi_2(x) \tag{5}$$

where $\varphi_1(x)$ is the wave function of slit 1, $\varphi_2(x)$ is the wave function of slit 2, and α and β are arbitrary complex numbers satisfying $|\alpha|^2 + |\beta|^2 = 1$.

$P(x) = |\varphi(x)|^2$ determines the probability (density) that a particle in state $\varphi(x)$ will be found at position x. $P_\alpha = |\alpha|^2$ is the probability of the photon passing through slit 1, and $P_\beta = |\beta|^2$ is the probability of the photon passing through slit 2. $f_1 (or f_2)$ is the curve observed by closing slit 2 (or slit 1). f_{12} is the curve observed by opening both slit 1 and slit 2. Therefore, the curves f_1, f_2 and f_{12} are measured as:

$$f_1 = |\alpha|^2 |\varphi_1(x)|^2 \tag{6}$$

$$f_2 = |\beta|^2 |\varphi_2(x)|^2 \tag{7}$$

$$\begin{aligned} f_{12}(x) &= \left|\varphi_p(x)\right|^2 = |\alpha\varphi_1(x) + \beta\varphi_2(x)|^2 \\ &= f_1 + f_2 + 2\sqrt{f_1 f_2} cos\theta \end{aligned} \tag{8}$$

where θ is the angle of the complex number $\alpha\varphi_1(x)\beta\varphi_2(x)$. $I = 2\sqrt{f_1 f_2} cos\theta$ is called the interference term. I is a necessary component of the quantum probabilistic model describing the distribution of the frequency of the photon detected by the detectors when both slits are open.

Quantum interference provides a comprehensive mathematical formalism to capture the intra-utterance interactions, which will be detailed in Section 4.3.2.

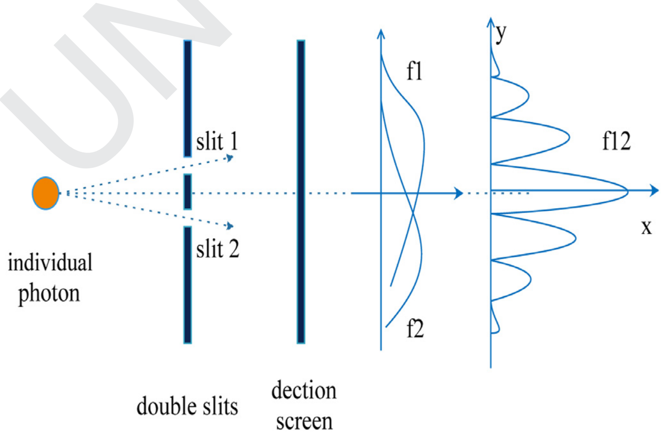

Fig. 2. The double-slit experiment. $f_1 (or f_2)$ is the curve observed by closing slit 2 (or slit 1). f_{12} is the curve observed by opening both slit 1 and slit 2. $f_{12} \neq f_1 + f_2$ because of the interference effect.

4. The quantum-like multimodal network framework

4.1. Problem formulation and overall framework

We target determining the attitude of each speaker at the utterance (sentence) level. The problem we investigate thus takes each utterance u as input and produces its sentiment label y as output. Hence, we formulate the problem as follows:

Given a multiturn conversation among speakers, how can we capture the interactions among them, and how can we determine their emotional changes brought by these interactions?

The architecture of the proposed quantum-like multimodal network (QMN) framework is shown in Fig. 3. We first extract textual and visual features for each utterance (turn) in the conversational discourse $x^{\vec{text}} = [\vec{r^t}_1, \vec{r^t}_2, \ldots, \vec{r^t}_n]$, $x^{\vec{img}} = [\vec{r^i}_1, \vec{r^i}_2, \ldots, \vec{r^i}_n]$, through a density matrix-based convolutional neural network (DM-CNN). Second, inspired by quantum measurement theory, a strong-weak influence model is developed to compute the inter-utterance influences among speakers within the whole conversation, denoted by R. Third, a variant of LSTM is built on top of the extracted multimodal features $x^{\vec{text}}$, $x^{\vec{img}}$ to model the evolution of sentiments in the conversation, with the output gate o_t combined with the inter-utterance influences R. Finally, inspired by quantum interference, we propose a multimodal decision fusion approach to obtain the completed sentiment decision (label) y_d. The details of these steps will be given in the next subsections.

4.2. Multimodal representation learning

Currently, a series of pioneering studies provide evidence that the density matrix, which is defined in the quantum probability space, could be applied in natural language processing as an effective representation method [13,25,27,29]. Compared with the embedding vector, the density matrix can encode 2-order semantic dependencies. Motivated by Zhang's work [29], we develop a density matrix-based convolutional neural network (DM-CNN) to represent the texts and images of all the

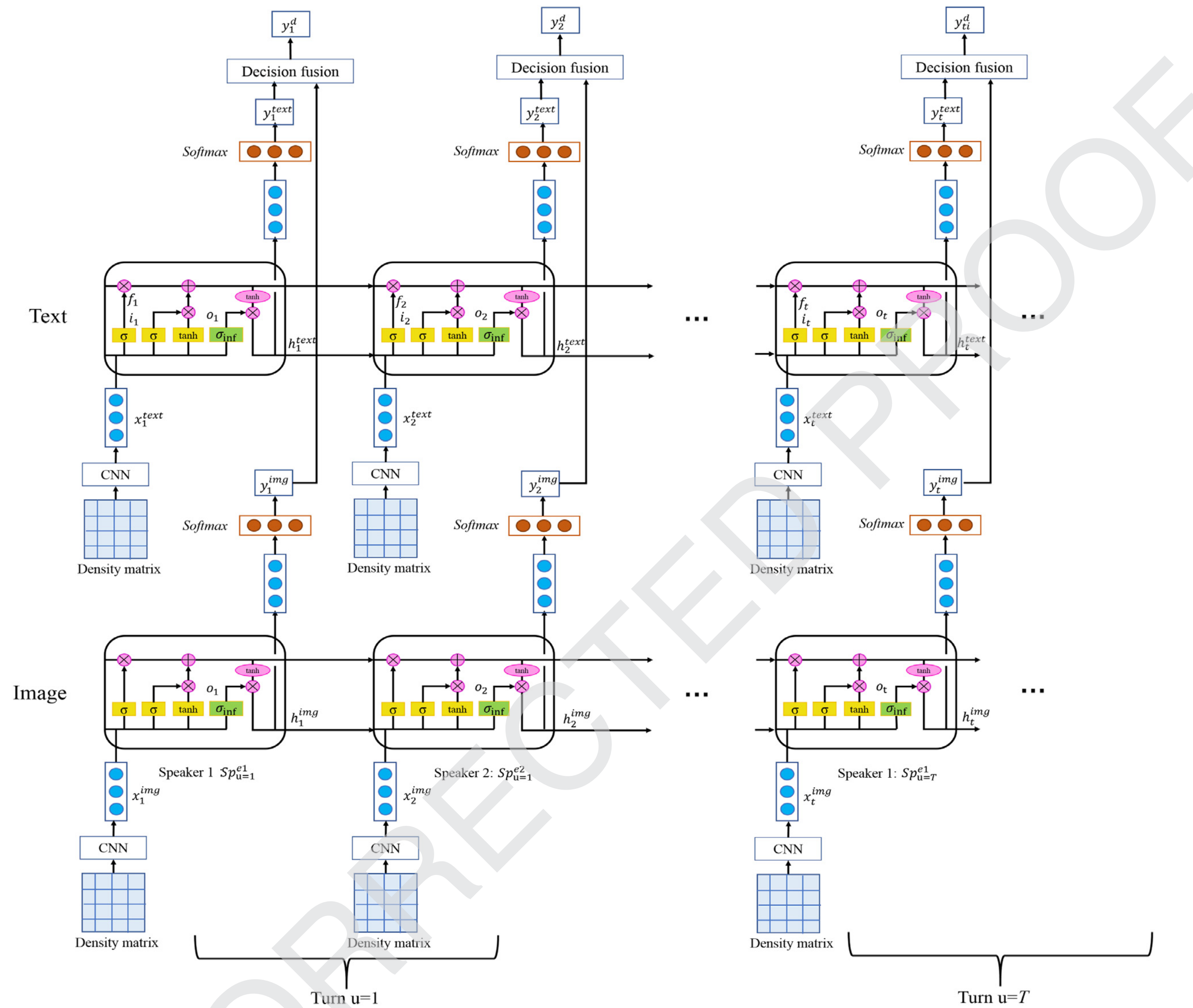

Fig. 3. The architecture of quantum-like multimodal network.

utterances in a conversation. The representation procedure for each modality is described below.

Text Representation. For text, suppose $|w_i\rangle = (w_{i1}, w_{i2}, \ldots, w_{id})^T$ is a normalized word vector. The projector Π_i for a single word w_i is formulated in Equation 9. The one-hot representation of words over other words is known to suffer from the curse of dimensionality and has difficulty representing ambiguous words. Therefore, we use word embeddings to construct projectors in semantic space. In this paper, we employ the GloVe tool [75] to find each word's embedding.

$$\Pi_i = |w_i\rangle\langle w_i| = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \ldots \\ w_{id} \end{pmatrix} \times (w_{i1}, w_{i2}, \ldots, w_{id}) = \begin{bmatrix} (w_{i1})^2 & w_{i1}w_{i2} & \ldots & w_{i1}w_{id} \\ w_{i2}w_{i1} & (w_{i2})^2 & \ldots & w_{i2}w_{id} \\ \vdots & & \ldots & \\ w_{id}w_{i1} & w_{id}w_{i2} & \ldots & (w_{id})^2 \end{bmatrix} \tag{9}$$

After defining the projector Π_i for each textual word, we represent a document (i.e., an utterance) with a density matrix, which can be formulated as:

$$\rho = \sum_i \Pi_i = \sum_i p_i |w_i\rangle\langle w_i| = \begin{bmatrix} \sum_i p_i (w_{i1})^2 & \sum_i p_i w_{i1} w_{i2} & \ldots & \sum_i p_i w_{i1} w_{id} \\ \sum_i p_i w_{i2} w_{i1} & \sum_i p_i (w_{i2})^2 & \ldots & \sum_i p_i w_{i2} w_{id} \\ \vdots & & \ldots & \\ \sum_i p_i w_{id} w_{i1} & \sum_i p_i w_{id} w_{i2} & \ldots & \sum_i p_i (w_{id})^2 \end{bmatrix} \tag{10}$$

where p_i is the corresponding probability of an event (word) Π_i, satisfying $\sum_i p_i = 1$. In quantum theory, how to calculate the probability of each quantum event has long been an open problem. In this work, we adopt one natural idea: to use the occurrence frequencies of words to compute their probabilities and the density matrix.

Now, we have obtained a density matrix ρ_t that temporarily represents the text part of the document. ρ_t is then fed into a deep CNN architecture to learn more abstract textual features, i.e., $x^{\vec{text}} = [\vec{r^t}_1, \vec{r^t}_2, \ldots, \vec{r^t}_n]$. The CNN consists of two convolutional layers, a fully connected layer and one softmax layer. Each convolutional layer is connected to a max pooling layer. The first convolutional layer has eight

5×5 filters. The second convolutional layer has sixteen 3×3 filters. The fully connected layer consists of 128 neurons. Note that the textual features $\vec{x^{text}}$ will be used as inputs for the QMN model.

Image Representation. We consider an image a document of visual words, in which each visual word is equivalent to a word in a text document. Therefore, we use these visual words $|s_i\rangle = (s_{i1}, s_{i2}, \ldots, s_{id})^T$ to construct visual projectors. The process of extracting visual words s_i is as described in the following procedure: (a) the SIFT features are extracted from all the images, and each SIFT feature is a 128-dimensional vector; (b) these extracted SIFT features are clustered to obtain k cluster centers through the k-means clustering algorithm. Each cluster center is a visual word, and all k visual words form a visual dictionary V, i.e., $V = \{s_1, s_2, \ldots, s_k\}$; (c) these visual words s_i are used to construct projectors $\Pi_i = |s_i\rangle\langle s_i|$ using Equation 9 and density matrices ρ_i using Equation 10.

Next, ρ_i is input into a deep CNN architecture. The image CNN is composed of six convolutional layers, one fully connected layer and one softmax layer. Each convolutional layer is connected to a max pooling layer. The first convolutional layer consists of 8 filters of size 7×7. The second convolutional layer consists of 16 filters of size 5×5. The third convolutional layer consists of 32 filters of size 5×5. The fourth convolutional layer consists of 64 filters of size 3×3. The fifth convolutional layer consists of 128 filters of size 3×3, and the sixth convolutional layer consists of 128 filters of size 2×2. This network is followed by the fully connected layer (size of 128) and the softmax layer. Finally, the activation values of the fully connected layer are used as the visual features for each utterance. The visual features $\vec{x^{img}}$ will be used as inputs for the QMN model.

4.3. Modeling interaction dynamics with the quantum-like multimodal network

In this subsection, we first propose a quantum measurement-inspired strong-weak influence model to capture the social influence among different speakers. Second, we introduce a quantum interference-inspired multimodal decision fusion approach to model the mutual influence between the text and image. Finally, we present the QMN model in detail.

4.3.1. Quantum measurement-Inspired strong-Weak influence model

Influence is an indirect, invisible way of altering the thought, behavior or nature of an entity, which is a difficult task to model [76]. When one talks to other people, he or she is influenced by the other people's styles of interaction. In a conversation, a speaker's affective state might or might not change, depending on the intensity of interaction. If the speaker's affective state changes, we argue that he or she is strongly affected by others. We call this a **strong interaction**. Similarly, if one speaker's words have a very small influence and lead to no changes to another speaker's affective state, we call this a **weak interaction**.

In QT, quantum measurement describes the interaction (coupling) between a quantum system and the measurement device. Strong measurement leads to the collapse of the quantum system state, while weak measurement disturbs the quantum system state very little. The variance in pointer readings of the measurement device could distinguish strong from weak interactions. In this work, we treat each speaker as a learning system. Accordingly, the interaction could be characterized as a coupling of two systems. The interaction between a quantum system and the measurement device is analogous to the interaction between two speakers. Some fundamental analogies exist between them in terms of the effect of the measurement/interaction. For example, both of them describe the interactions of different strengths between the two systems. Strong measurement involves a change from a superposition state to the eigenstate, while strong interaction also makes a change from the original affective state to another affective state. On the other hand, weak measurement and weak interaction can hardly disturb the system/affective state. Therefore, quantum measurement provides us with natural inspiration and rigorous mathematical formalism to help understand and model complex interactions among speakers; we model strong and weak interactions with the formalism of quantum measurement and thus develop a strong-weak influence model.

Specifically, we base our strong-weak influence model on the dynamic "influence model", which is a generalization of HMMs for describing the influence that each Markov chain has on the others through constructing influence matrices [76]. This model gives an abstract definition of influence: an entity's state is influenced by its neighbors' states and changes accordingly. Each entity has an influence on every other entity in the network.

Dynamic influence model

Suppose there are C entities in the system, and each entity e is associated with a finite set of possible states $\{1, 2, \ldots, S\}$. Note that to avoid confusion between the *time* in the influence model and that in LSTM, we use $\boldsymbol{u}$ to represent the time series (turn) in the influence model and use $\boldsymbol{t}$ to denote each time step in the LSTM networks.

At each different turn u, each entity e is in one of the states, denoted by $q_u^e \in \{1, 2, \ldots, S\}$. Each entity emits an observable o_u^e at turn u following the emission probability $b_{q_u^e}(o_u^e) = P(o_u^e|q_u^e)$. Influence is treated as the conditional dependence among each entity's current state q_u^e at turn u and the previous states of all the entities $q_{u-1}^1, q_{u-1}^2, \ldots, q_{u-1}^C$ at turn $u-1$. Apparently, q_u^e is only influenced by all entities at turn $u-1$. Therefore, the conditional probability can be formulated as:

$$P\left(q_u^e|q_{u-1}^1, q_{u-1}^2, \ldots, q_{u-1}^e, \ldots, q_{u-1}^C\right) = \sum_{c\in 1,2,\ldots,C} R(r_u)_{e,c} \times Infl\left(q_u^e|q_{u-1}^c\right) \quad (11)$$

where $R(r_u)$ is a $C\times C$ matrix and $R(r_u)_{e,c}$ represents the element in the eth row and the cth column; $r_u \in \{1, 2, 3, \ldots, J\}$, $u = 1, \ldots, T$; and J is a hyperparameter set freely by the user to define the number of influence matrices $R(r_u)$ for improving the adaptability of the influence model. $Infl\left(q_u^e|q_{u-1}^c\right)$ is modeled using an $S\times S$ matrix $M^{c,e}$, namely, $Infl\left(q_u^e|q_{u-1}^c\right) = M^{c,e}_{q_{u-1}^c, q_u^e}$, where $M^{c,e}_{q_{u-1}^c, q_u^e}$ represents the element in the q_{u-1}^cth row and q_u^eth column of matrix $M^{c,e}$. The matrix $M^{c,e}$ is similar to the transition matrix, which can be simplified by two $S\times S$ matrices: E^c and F^c. E^c captures the self-state transition, i.e., $E^c = M^{c,c}$, and F^c represents the adjacent state transition, i.e., $F^c = M^{c,e}, \forall e \neq c$.

Quantum-Inspired Strong-Weak Influence Model

However, in a turn-taking conversation, only the first speaker's state at each turn, denoted by $q_u^e|_{e=1}$, is influenced by the previous states of all the entities, while the remaining speakers' states at each turn, denoted by $q_u^e|_{e\geq 2}$, are influenced by both the current states of the speakers who speak in front of e at turn u, i.e., $q_u^1, q_u^2, \ldots, q_u^{e-1}$, and the previous states of the other speakers who have not yet spoken (including the current speaker under concern) in the current round, i.e., $q_{u-1}^e, q_{u-1}^{e+1}, \ldots, q_{u-1}^C$. Then, the conditional probability is divided into two parts:

$$\begin{cases} P\left(q_u^e, e = 1|q_{u-1}^1, q_{u-1}^2, \ldots, q_{u-1}^C\right) \\ P\left(q_u^e, e \geq 2|q_u^1, q_u^2, \ldots, q_u^{e-1}, q_{u-1}^e, q_{u-1}^{e+1}, \ldots, q_{u-1}^C\right) \end{cases} \quad (12)$$

Referring to the example shown in Fig. 1, we have $C = \{Jen(J), Ross(R)\}$. Each speaker is in one of three affective states, which are positive, negative and neutral; i.e., $S = 3$, and $q_u^R, q_u^J \in \{-1, 0, 1\}$. Hence, speaker *Jen*'s affective state q_u^J at turn u is influenced by the previous states of both J and R at turn $u-1$, i.e., q_{u-1}^J, q_{u-1}^R. *Ross*'s affective state q_u^R is influenced by both his own previous state q_{u-1}^R at turn $u-1$ and *Jen*'s state in the current turn q_u^J. The conditional probability is measured as:

$$\begin{cases} P\left(q_u^J|q_{u-1}^J, q_{u-1}^R\right) \\ = R(r_u)_{JJ} \cdot Infl(q_u^J|q_{u-1}^J) + R(r_u)_{JR} \cdot Infl(q_u^J|q_{u-1}^R) \\ P\left(q_u^R|q_u^J, q_{u-1}^R\right) \\ = R(r_u)_{RJ} \cdot Infl(q_u^R|q_u^J) + R(r_u)_{RR} \cdot Infl(q_u^R|q_{u-1}^R) \end{cases} \quad (13)$$

where $R(r_u)_{JJ}$, $R(r_u)_{JR}$, $R(r_u)_{RJ}$, and $R(r_u)_{RR}$ are four elements of the influence matrix $R(r_u)$. Each element is also a 3×3 matrix, which de-

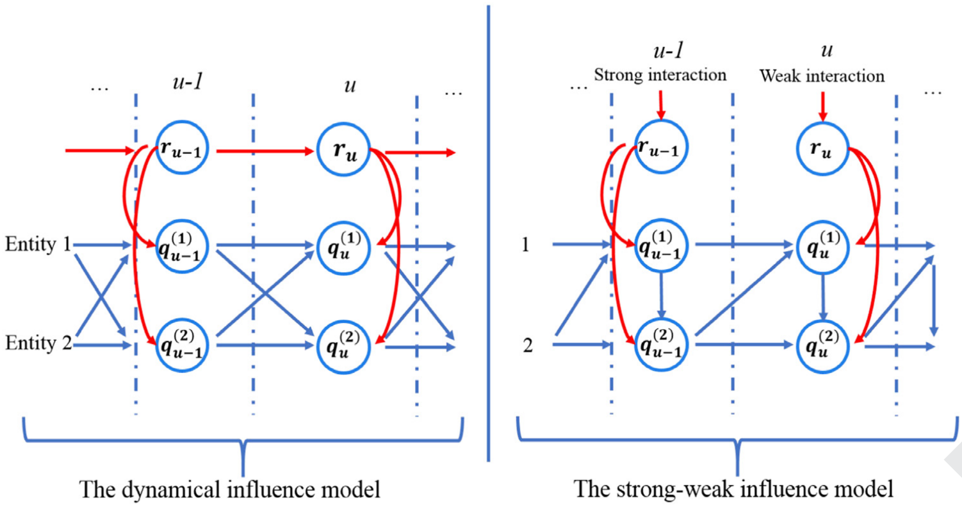

Fig. 4. The difference between the dynamic influence model and the strong-weak influence model. The blue lines show the dependence, and the red lines indicate the switching capacity of the influence model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

notes, in different affective states (-1, 0, 1), how *Jen* influences herself, how *Ross* influences *Jen*, how *Jen* influences *Ross*, and how *Ross* influences himself, respectively. $Infl(q_u^J|q_{tu-1}^J)$, $Infl(q_u^J|q_{u-1}^R)$, $Infl(q_u^R|q_u^J)$, and $Infl(q_u^R|q_{u-1}^R)$ are four 3×3 transition matrices.

Inspired by quantum measurement, we use two influence matrices (i.e., $J=2, r_u \in \{1,2\}$) to represent strong and weak influences. The switching of r_u is determined by the average standard deviation of the speakers' sentimental scores σ_{avg}. We set the eigenvalues of the speaker's affective state to -1, 0 and 1; i.e., $x \in \{-1, 0, 1\}$. Hence, we introduce the following prior for r_u:

$$\begin{cases} r_t = 1 & \text{if } \sigma_{avg} \geq \sum_x p(x)|x| \; weak\,influence \\ r_t = 2 & \text{if } \sigma_{avg} < \sum_x p(x)|x| \; strong\,influence \end{cases} \tag{14}$$

where $p(x) = \left(2\sigma^2\pi\right)^{-\frac{1}{2}} e^{-(x-\mu_{avg})^2/2\sigma^2}$, denoting the probability amplitude to obtain x, and in this work, μ_{avg} is set to the average of all expectations.

We illustrate the difference between the dynamic influence model and the strong-weak influence model in Fig. 4. Finally, we obtain two influence matrices, which capture the strong and weak influences of one speaker on another speaker under different interactive environments. The detailed inference process is given in Appendix A.

4.3.2. A Quantum Interference-Inspired multimodal decision fusion approach

In the process of identifying the overall sentiment of multimodal content, a user commonly makes a decision simultaneously based on his or her understanding of the content through multiple channels corresponding to different modalities, which could cause cognitive interference. Note that in this paper, the user's cognitive state mainly refers to the user's state of mind that determines his or her judgment about the sentiment of an utterance, and it is involved in conversations influenced by the previous utterances. The judgment result may be biased to the positive or negative polarity variations. Before a user reads the text and sees the image, the user's cognitive state is a superposition of the sentiments of multimodalities, which means that his or her cognitive state is uncertain and indefinite. Note that a user's cognitive state mainly refers to his or her cognition and judgment of emotions in this paper. In such a superposition-like state, he or she does not make a specific decision on the sentiment category of multimedia content. After he or she reads the text and sees the image, his or her cognitive state may collapse to one of the sentiment scores (+2, +1, -1, -2).

We draw an analogy to the double-slit experiment in multimodal sentiment analysis. The original decision result is uncertain, which can be considered as the photon. The sentiment in the text and the image can be seen as two slits, and each sentiment score is a position on the detection screen, as shown in Fig. 5. In our analogy, the decision result is in a superposition-like state for the sentiment of the text and the image, so that the sentiment information of each modality will simultaneously influence the final decision. Note that we elaborate on this analogy for developing a new fusion approach instead of modeling the quantum process.

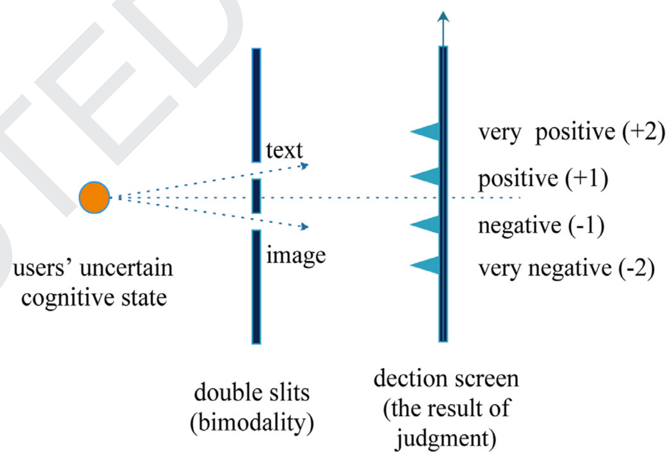

Fig. 5. Our double-slit experiment analogy for multimodal sentiment analysis.

We use the wave function $\varphi(x)$ to formalize our analogy. The decision result is in a superposition of the sentiment of the text and the image, as shown below:

$$\varphi_d(x) = \alpha\varphi_t(x) + \beta\varphi_i(x) \tag{15}$$

where $\varphi_t(x)$ and $\varphi_i(x)$ are the wave functions of the sentiment of the text and the image, respectively. Therefore, the probability distribution of making decisions only through the text or the image can be formulated as:

$$\begin{aligned} f_t &= |\alpha|^2 |\varphi_t(x)|^2 \\ f_i &= |\beta|^2 |\varphi_i(x)|^2 \end{aligned} \tag{16}$$

The probability distribution of the final decision can be measured as:

$$\begin{aligned} f_d(x) &= |\varphi_d(x)|^2 = |\alpha\varphi_t(x) + \beta\varphi_i(x)|^2 \\ &= |\alpha\varphi_t(x)|^2 + |\beta\varphi_i(x)|^2 + 2|\alpha\varphi_t(x)\beta\varphi_i(x)|cos\theta \\ &= f_t + f_i + 2\sqrt{f_t f_i}cos\theta \end{aligned} \tag{17}$$

At the decision level, we interpret $P_t(x) = |\varphi_t(x)|^2$ as the probability that the sentiment score of the text is x, denoted by P_t. We interpret $P_i(x) = |\varphi_i(x)|^2$ as the probability that the sentiment score of the image is x, denoted by P_i. The final decision P_d can be written as:

$$P_d = \alpha^2 P_t + \beta^2 P_i + 2\alpha\beta\sqrt{P_t P_i}cos\theta \tag{18}$$

where α^2 and β^2 are the normalized weights assigned to the text and the image decision, respectively. $I = 2\alpha\beta\sqrt{P_t P_i}cos\theta$ is the interference term, which represents the degree to which local decisions conflict.

4.3.3. Quantum-like multimodal network

In Sections 4.3.1 and 4.3.2, we covered the interaction information (including interactions between modalities and those among speakers); next, we can incorporate them into the quantum-like multimodal network (QMN), which is detailed in this subsection.

Here, we first briefly review the standard LSTM network to establish the basis for understanding the proposed QMN model. The long short-term memory (LSTM) network, a special kind of gated RNN, was introduced by Hochreiter and Schmidhuber [77]. A common architecture of the LSTM network is composed of a memory cell, a forget gate, an input gate, and an output gate. The memory cell flows straight down the entire chain, storing information for either long or short time periods. The forget gate determines what information to discard in the cell. The input gate controls what new information would be stored in the cell. The output gate controls the output value of the LSTM unit based on the memory cell. Specifically, LSTM is written as below:

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{19}$$

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{20}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{21}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{22}$$

$$h_t = o_t \odot tanh\left(c_t\right) \tag{23}$$

where f_t, i_t, o_t, and c_t are the forget gate, input gate, output gate and cell state at time t, respectively. $W_{xf}, W_{hf}, \ldots, W_{hc} \in \mathbb{R}^{d\times d}$ are the weighted matrices, and $b_f, b_i, b_o \in \mathbb{R}^d$ are biases to be learned during training. σ is the sigmoid function, *tanh* is the hyperbolic tangent function, and $\odot$ denotes pointwise multiplication. x_t and h_t represent the inputs and outputs, respectively.

Then, as a modification to the standard LSTM, the QMN model is composed of two parts, which can model the text and image interactively. The main idea is (1) for each LSTM unit, the output gate o_t is combined with the learned influence matrices **R** to constitute a new output gate, describing what information we are going to output. Thus, the new output gate explicitly considers the previous speakers' influences. (2) Taking the textual and visual vectors built by the DM-CNN as inputs, their hidden states h_{text}, h_{img} are obtained using the extended LSTM networks. (3) With this design, the QMN model makes local decisions on the text and the image and fuses them at the decision level using the quantum interference-inspired multimodal fusion approach, which is detailed in Section 4.3.2. Fig. 3 depicts the overall architecture of the QMN model.

Let us first formalize the notation. x_t^{text}, x_t^{img} and h_t^{text}, h_t^{img} represent the inputs and outputs of each LSTM unit t of the text and image, where $t = \{1, 2, \ldots, N\}$, N is the number of speaker utterances. $x^{\vec{text}} = [\vec{r^t}_1, \vec{r^t}_2, \ldots, \vec{r^t}_n]$, and $x^{\vec{img}} = [\vec{r^i}_1, \vec{r^i}_2, \ldots, \vec{r^i}_n]$ are the vector representations of the text and image, which are learned by the DM-CNN, and h_t^{text}, h_t^{img} are considered the output feature representations of multimodal utterances. Since our aim is to identify the sentiment polarity of each utterance, we first put h_t^{text}, h_t^{img} into the *softmax* layer to obtain the probability decisions of the sentiment label y_t^{text}, y_t^{img} and thus merge them to yield the final decision y_t^d. That is,

$$y_t^S = softmax\left(W_s h_t^S + b_s\right) \tag{24}$$

$$y_t^d = \alpha^2 y_t^{text} + \beta^2 y_t^{img} + 2\alpha\beta\sqrt{y_t^{text} y_t^{img}}cos\theta \tag{25}$$

where $S \in \{text, img\}$, W_s and b_s are the parameters for the *softmax* layer. α^2 and β^2 are the normalized weights assigned to the text and the image decision. $I = 2\alpha\beta\sqrt{y_t^{text} y_t^{img}}cos\theta$ is the interference term, which represents the degree of conflicting local decisions.

In a conversation, the influence that one speaker has on another controls the affected speaker's response. In Fig. 3, for two adjacent speakers (denoted by $e1$ and $e2$) at turn $u = 1$ (i.e., $Sp^{e1}_{u=1}$, $Sp^{e2}_{u=1}$), $Sp^{e1}_{u=1}$ actually determines how $Sp^{e2}_{u=1}$ is constructed. Furthermore, at the next turn $u = 2$, the construction of $Sp^{e1}_{u=2}$ is influenced by both $Sp^{e1}_{u=1}$ and $Sp^{e2}_{u=1}$, and the construction of $Sp^{e2}_{u=2}$ is influenced by both $Sp^{e2}_{u=1}$ and $Sp^{e1}_{u=2}$. Influence controls what information one speaker is going to output, which is similar to the role of the output gate in the LSTM network. This influence has already been described by the influence matrix **R** (subsection 4.3.1). Hence, we consider the influences on the next speaker from the previous speakers by incorporating the influence scores into the *sigmoid* function in the quantum-like multimodal network, which can be formulated as:

$$\begin{aligned}
o^{e1}_{u|u=1} &= \sigma\left(W_{xo}\vec{x}^{e1}_u + b_o\right) \\
o^{e2}_{u|u=1} &= \sigma\left(W_{xo}\vec{x}^{e2}_u + W_{ho}h^{e1}_u + b_o\right) + \sigma(R_{e2,e1} \cdot \vec{x}^{e2}_u) \\
o^{e1}_{u|u\geq 2} &= \sigma\left(W_{xo}\vec{x}^{e1}_u + W_{ho}h^{e2}_{u-1} + b_o\right) + \sigma(W_{e1}[R_{e1,e1}, R_{e1,e2}] \cdot \vec{x}^{e1}_u) \\
o^{e2}_{u|u\geq 2} &= \sigma\left(W_{xo}\vec{x}^{e2}_u + W_{ho}h^{e1}_u + b_o\right) + \sigma(W_{e2}[R_{e2,e2}, R_{e2,e1}] \cdot \vec{x}^{e2}_u)
\end{aligned} \tag{26}$$

where $u = 1, 2, \ldots, T,$ denotes the number of turns and W_{e1} and W_{e2} are the normalized weights. $R_{e1,e1}$, $R_{e1,e2}$, $R_{e2,e1}$, and $R_{e2,e2}$ are elements in the influence matrices $R(r_u)$.

Model Training. In the QMN model, we need to optimize all the parameters, denoted by Θ: [W_{xi}, W_{hi}, b_i, W_{xf}, W_{hf}, b_f, W_{xo}, W_{ho}, b_o, W_{e1}, W_{e2}, W_{xc}, W_{hc}, b_c, W_s, b_s]. Cross-entropy with L_2 regularization is used as the loss function, which is defined as:

$$J = -\frac{1}{N}\sum_i\sum_j y_i^j log\hat{y}_i^j + \lambda_r\|\theta\|^2 \tag{27}$$

where y_i denotes the ground truth and $\hat{y}_i$ is the predicted sentiment distribution. i is the utterance index, and j is the class index. λ_r is the coefficient for $L2$ regularization.

We use the backpropagation method to compute the gradients and update all the parameters Θ by:

$$\Theta = \Theta - \lambda_l \frac{\partial J(\Theta)}{\partial \Theta} \tag{28}$$

where λ_l is the learning rate. To avoid overfitting, we use a dropout strategy to randomly omit half of the feature detectors in each training case.

5. Experiments

5.1. Experimental settings

Our main research questions are as follows: (1) Is the interaction information important in conversational sentiment analysis? (2) How can the influence of one speaker on another be presented? (3) Which component of the QMN plays a key role in the performance?

To answer (1), we compare the performance of the QMN model with a number of baselines and study the importance of modeling interactions. To answer (2), we visualize the influence matrices and conduct a detailed analysis. To answer (3), we conduct an ablation test by adopting only one component at a time and evaluate their impacts on the overall performance.

Table 2
Sentiment distributions in the MELD and IEMOCAP datasets.

Dataset	Sentiment Category	No. of Utterances		
		Train	Dev	Test
MELD	positive	2334	233	521
	neutral	4710	470	1256
	negative	2945	406	833
	anger	1109	153	345
	disgust	271	22	68
	fear	268	40	50
	joy	1743	163	402
	neutral	4710	470	1256
	sadness	683	111	208
	surprise	1205	150	281
IEMOCAP	anger	804	#	158
	happiness	377	#	127
	sadness	592	#	191
	neutral	1124	#	357
	other	3600	#	1092

Datasets. Given that multimodal sentiment analysis of conversations is a new area, the benchmark datasets are relatively limited. In this work, we perform experiments on the MELD[1] [22] and IEMOCAP[2] datasets [78]. MELD contains 13,708 utterances from 1433 dialogues of Friends TV series. The utterances in each dialogue are annotated with one of three sentiments (positive, negative of neutral) and one of seven emotions (anger, disgust, fear, joy, neutral, sadness or surprise). The utterances in MELD are multimodal, encompassing audio and visual as well as textual information. In this work, we only use textual and visual information.

IEMOCAP is a multimodal database of ten speakers involved in two-way dyadic conversations. Each utterance is annotated using one of the following emotion categories: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, or others. We consider the first four categories and assign other emotions to the fifth category to compare our QMN with other state-of-the-art baselines in a fair manner.

The details about the training/development/testing split are provided in Table 2, which also provides the sentiment distribution information for all the datasets.

Preprocessing. The multimodal data are preprocessed as follows. For the image information, the overly large images (i.e., size exceeding 1000 pixels*1000 pixels) are re-sized to 360*640. For textual information, we first clean all the texts by checking for illegible characters and correcting spelling mistakes automatically. The stop words are removed using a standard stopword list from Python's NLTK package [79]. We do not filter out the punctuation marks since some punctuation marks, such as question marks and exclamation points, tend to carry subjective information. We run the experiments using five-fold cross-validation on all the comparative models.

Evaluation metrics. Since our approach and baselines are supervised sentiment analysis methods, we adopt the **precision**, **recall**, **F1 score**, **andaccuracy** as the evaluation metrics to evaluate the classification performance of each method. Note that considering the imbalanced sample problem, we adopt the weighted F1 score and set *class-weight* to "*balanced*" during the training process. We employ the paired *t*-test to perform significance test and report the standard errors of the difference between the means (denoted by *sed*) in Table 3 and Table 4.

Hyperparameter Setting. In this work, we use the GloVe word vectors[3] [75] to produce quantum projectors. The dimensionality of the embeddings is set to 300. Similarly, we set the dimensionality of the visual words to 128, which is the default setting of the SIFT algorithm. All weight matrices are given their initial values by sampling from a uniform distribution $U(-0.1, 0.1)$, and all biases are set to zero. We use the Adam [80] algorithm to train the network, and the best learning rate is set to 0.002 for the 3-class and 5-class classification tasks and 0.005 for the 7-class classification task. The batch size is 60. TensorFlow [81] is used for implementing our neural network models. The coefficient of $L2$ normalization in the objective function is set to 10^{-5}, the number of epochs is set to 50, and the dropout rate is set to 0.5.

5.2. Comparative models

To verify the effectiveness of the proposed QMN model, we compare our model with a number of baselines. They are listed as follows.

(1) **Single textual model:** we apply a deep convolutional neural network (CNN) on each utterance to extract the textual features and feed these features into the softmax classifier to predict the sentiment. The CNN includes three convolutional layers, three pooling layers, and a fully connected layer. The first convolutional layer has eight 5×5 filters. The second convolutional layer has sixteen 3×3 filters. The third convolutional layer has thirty-two 2×2 filters. We set the batch size to 60 and the dimensionality of the word embeddings to 300.
(2) **Single visual model:** an image sentiment prediction framework is built with a convolutional neural network (CNN). A deep CNN architecture is used for learning visual features and predicting the sentiment. The CNN includes six convolutional layers and one fully connected layer. The setting of filters is consistent with the abovementioned density matrix-based CNN. We set the batch size to 60 and the dimensionality of the visual features to 128.
(3) **Feature-level multimodal fusion (FMF) model:** the FMF model takes joint text-level and image-level representations as input, and two kinds of representations are extracted by a single textual model and a single visual model. Then, the FMF model trains a logistic regression classifier (whose parameters are set to the default values) to identify the sentiment polarities of multimodal documents.
(4) **Dempster-Shafer evidence fusion (DSEF) model:** as a mathematical theory of evidence, the Dempster-Shafer (D-S) evidence theory allows one to combine evidence from different sources and arrive at a degree of belief that takes into account all the available evidence [82]. In this paper, a single textual model and a single visual model return two results lists with different probability scores. Hence, three (or more) sentiment scores (which are 0, +1, and -1) construct the power set. We use the probability scores to specify the mass function. According to the D-S evidence theory, the combination (called the joint mass) is calculated from the two sets of masses m_{text} and m_{image} in the following manner: $m_{multimodal}(A) = (m_{text} \oplus m_{image})(A) = \frac{1}{1-k}\sum_{B\cap C=A} m_{text}(B)m_{image}(C)$, where $K = \sum_{B\cap C=\varnothing} m_{text}(B)m_{image}(C)$.
(5) **Multimodal deep learning (MDL) model:** this model can learn a joint representation of various features extracted in different modalities, which is similar to the method proposed in [83]. In [83], the authors used a restricted Boltzmann machine (RBM) to learn the joint distribution over image and text inputs. We choose to replace the RBM with a convolutional neural network (CNN) to learn the joint distribution over our image and text inputs by constructing a shared hidden layer based on a similar framework.
(6) **Convolutional recurrent neural network (CRNN)**: a CRNN [84] designs a hybrid deep learning structure that integrates a convolutional neural network (CNN) and a recurrent neural network (RNN) for conducting emotion recognition tasks in one single framework. Specifically, the CNN is used for learning textual and visual features and mining the intermodality correlation

[1] https://affective-meld.github.io/.
[2] http://sail.usc.edu/iemocap/.
[3] Pretrained word embeddings for GloVe can downloaded from https://nlp.stanford.edu/projects/glove/.

Table 3
Performance of all the baselines on the MELD dataset. The best-performing system is indicated in bold. The numbers in parentheses indicate the relative improvement achieved by our QMN model over the hierarchical contextual LSTM network model, which appears to be the best-performing model among the comparative models. The symbol † indicates statistically significant improvement over all the baselines.

MELD dataset	Model	Evaluation metric			
		Precision	Recall	F1	Accuracy
Sentiments (3-class)	Single textual model	0.601	0.621	0.584	0.621
	Single visual model	0.412	0.434	0.426	0.434
	FMF model	0.516	0.533	0.509	0.533
	DSEF model	0.449	0.519	0.457	0.519
	MDL model	0.556	0.571	0.563	0.572
	CRNN model	0.619	0.566	0.571	0.577
	Contextual h-LSTM network	0.684	0.693	0.675	0.693
	Hierarchical contextual h-LSTM network	0.695	0.707	0.693	0.707
	QMSA framework	0.644	0.659	0.653	0.659
	Textual DM-CNN model	0.651	0.669	0.657	0.668
	Visual DM-CNN model	0.447	0.471	0.459	0.471
	DM-QIMF model	0.700	0.719	0.704	0.720
	QMN model	**0.742**† (+6.76%) (*sed*: 0.0096)	**0.755** † (+6.79%) (*sed*: 0.0104)	**0.729**† (+5.19%) (*sed*: 0.0084)	**0.756**† (+6.79%) (*sed*: 0.0103)
Emotions (7-class)	Single textual model	0.520	0.558	0.532	0.558
	Single visual model	0.381	0.403	0.393	0.397
	FMF model	0.391	0.487	0.403	0.487
	DSEF model	0.473	0.500	0.480	0.501
	MDL model	0.332	0.481	0.392	0.481
	CRNN model	0.520	0.546	0.516	0.546
	Contextual h-LSTM network	0.575	0.645	0.584	0.645
	Hierarchical contextual h-LSTM network	**0.625**	0.664	0.615	0.664
	QMSA framework	0.581	0.640	0.612	0.640
	Textual DM-CNN model	0.569	0.608	0.572	0.607
	Visual DM-CNNmodel	0.395	0.417	0.408	0.417
	DM-QIMF model	0.587	0.660	0.617	0.659
	QMN model	0.552 (−11.68%) (*sed*: 0.0057)	**0.693**† (+4.15%) (*sed*: 0.0063)	**0.627**† (+1.96%) (*sed*: 0.0068)	**0.693**† (+4.15%) (*sed*: 0.0063)

Table 4
Performance of all the baselines on the IEMOCAP dataset. The best-performing system is indicated in bold. The numbers in parentheses indicate the relative improvements over the hierarchical contextual LSTM network model. The symbol † indicates statistically significant improvement over all the baselines.

IEMOCAP dataset	Model	Evaluation metric			
		Precision	Recall	F1	Accuracy
Sentiments (5-class)	Single textual model	0.534	0.564	0.538	0.564
	Single visual model	0.421	0.533	0.448	0.533
	FMF model	0.518	0.546	0.521	0.546
	DSEF model	0.563	0.570	0.567	0.570
	MDL model	0.322	0.567	0.411	0.567
	CRNN model	0.555	0.574	0.533	0.574
	Contextual h-LSTM network	0.600	0.615	0.590	0.618
	Hierarchical contextual h-LSTM network	0.609	0.625	0.602	0.625
	QMSA framework	0.570	0.595	0.574	0.595
	Textual DM-CNN model	0.556	0.590	0.563	0.589
	Visual DM-CNN model	0.446	0.554	0.470	0.554
	DM-QIMF model	0.592	0.628	0.603	0.628
	QMN model	**0.631**† (+3.61%) (*sed*: 0.0056)	**0.647**† (+3.68%) (*sed*: 0.0089)	**0.623**† (+3.49%) (*sed*: 0.0083)	**0.648**† (+3.68%) (*sed*: 0.0090)

through designed convolutional filters. The RNN is used to model the evolution, transition and long-term dependencies of the features for final sentiment prediction.

(7) **Contextual h-LSTM & hierarchical contextual h-LSTM network models**: we implement a contextual h-LSTM [20] network to model the semantic dependency among the utterances. Context-independent unimodal features, which are extracted by the CNN, are fed to the proposed h-LSTM network to obtain context-sensitive unimodal feature representations and sentiment labels for each utterance. Furthermore, we have also implemented a hierarchical deep network that consists of two levels: (1) context-independent unimodal features are fed to the proposed h-LSTM network to obtain context-sensitive unimodal feature representations for each utterance; (2) outputs from each h-LSTM network in (1) are concatenated and fed into the h-LSTM network, thus providing an inherent fusion scheme.

(8) **QMSA framework:** the QMSA [13] framework first adopts the quantum-inspired multimodal representation (QMR) model to represent the images and the texts separately and obtains their own local decisions using an RF classifier. Second, it fuses their decisions at the decision level to obtain the final results.

A series of our proposed submodels are listed below:

(9) **Textual DM-CNN:** a density matrix-based CNN architecture is used for learning textual features and predicting the sentiment of each utterance. CNN includes three convolutional layers, three pooling layers, and a fully connected layer. We set the learning rate to 0.002, the batch size to 60 and the dimensionality of word embeddings to 300.

label text

(10) **Visual DM-CNN:** we apply a density matrix-based CNN on each image to extract visual features and feed these features into the softmax classifier to predict the sentiment of each utterance. The CNN includes six convolutional layers and one fully connected layer. We set the learning rate to 0.002 and the batch size to 60.

(11) **DM-QIMF:** the DM-QIMF first adopts the textual and visual DM-CNN models to represent the texts and images separately and obtains their local decisions. Then, it fuses their decisions at the decision level to obtain the final results using the quantum interference-inspired fusion method.

5.3. Results on the MELD dataset

The first set of experiments is conducted on the MELD dataset, which generally provides more training samples of multimodal documents than the IEMOCAP dataset. The experimental results are summarized in Table 3, from which we can observe the following.

(1) In the case of sentiment classification, the single visual model performs poorly. This result indicates that it is insufficient to only utilize visual features to analyze the sentiment polarity of images. Compared with the single visual model, the single textual model improves performance, as we expected, because visual sentiment analysis involves a higher level of abstraction and subjectivity than textual sentiment. By concatenating textual features and visual features, the FMF model outperforms the single visual model but is outperformed by the single textual model. This finding shows that a simple concatenation strategy is not able to capture the correlation between multimodalities. As a general framework for reasoning with uncertainty, the Dempster-Shafer (D-S) evidence theory is also taken as a baseline. It achieves lower performance metrics than the FMF model. A reason is that this baseline largely relies on how to define the mass function and the judgment rule. As one of the earliest deep learning-based multimodal sentiment analysis methods, the MDL model outperforms the FMF and DSEF models, showing that learning a joint representation helps improve performance. The CRNN employs CNNs to extract multimodal features, puts them into an RNN structure, and achieves better classification scores than other models. An explanation is that the RNN effectively takes into account the sequence information, i.e., the sequence of the utterances.

Furthermore, by treating surrounding utterances as the context of the utterance to be classified, two different frameworks, the contextual and hierarchical h-LSTM frameworks, perform quite well on the MELD dataset. They achieve accuracy results of 69.3% and 70.7%, respectively, which are much higher than those of the other baselines. The reason is that they can effectively preserve the sequential order of utterances and enable consecutive utterances to share information. Through using the quantum-inspired representation, the QMSA framework outperforms the CRNN and MDL models, suggesting that an effective semantic learning model could help the machine to better "understand" multimodal documents. However, it is outperformed by the contextual and hierarchical contextual h-LSTM network models, probably because it does not model the inter-utterance dependencies.

Finally, compared with the single textual and visual models, the accuracy results achieved by the textual and visual DM-CNN models increased by 7.57% and 8.78%, respectively. Based on the calculations, there are approximately 685 textual utterances misclassified by a single textual model, while they have been accurately recognized by the textual DM-CNN model. There are approximately 792 visual utterances in MELD misclassified by a single visual model, while they are accurately recognized by the visual DM-CNN model. The textual and the visual DM-CNN models outperform both the single textual and visual models, which shows the effectiveness of the proposed density matrix representation method. By fusing their local results using our quantum interference-inspired multimodal fusion method, the DM-QIMF model performs very well, achieving the second highest experimental performance. Taking a further step towards emphasizing the importance of modeling interactions, the proposed QMN model achieves the best classification results on all metrics and significantly outperforms all the baselines. Compared with the nonhierarchical and hierarchical contextual h-LSTM network models, the accuracy results increased by 9.1% and 6.7%. Overall, we attribute the main improvements to both the quantum interference-inspired fusion strategy and the quantum measurement-inspired strong-weak influence model, which ensures that the QMN model can learn both intra- and inter-utterance interactions. A detailed ablation study is provided in Section 5.6.

(2) In the case of emotion classification, overall, we can observe that the performance of all the models has been reduced because of the increase in the number of classes. Nevertheless, we can still observe similar results. For example, the single textual model can achieve a higher F1 score and accuracy than the single visual model, which performs the worst. These results indicate that sentiment recognition from images is not as effective as that from text. The textual DM-CNN and the visual DM-CNN outperform both the single textual and visual models, showing the effectiveness of the proposed representation method. The FMF and DSEF models achieve poor performance in classifying the seven emotions. We notice that the performance of the MDL model declines sharply. The CRNN model outperforms the MDL model, which implies that distinguishing fine-grained emotions might be dependent on sequence information. The contextual and hierarchical contextual h-LSTM networks outperform the CRNN model in the MELD dataset by a margin of 17% to 22%. These results prove that modeling contextual dependencies among utterances improves the classification results. Our QMN model still achieves the best performance. Compared with the contextual and hierarchical contextual h-LSTM network models, the QMN model improves the performance by 7.4% and 4.2%, respectively. The main reason is that the QMN model models second-order semantic dependencies, previous speakers' influence and intra-correlations between modalities. The results demonstrate the effectiveness and necessity of modeling the interactions in conversational sentiment analysis. Furthermore, quantum probability theory has been proven to be an effective mathematical formalism to model complex interactions.

5.4. Results on the IEMOCAP dataset

Table 4 shows the performance comparison of the QMN model with the baselines on the IEMOCAP dataset, which is another widely used dyad conversational emotion dataset. Compared with the MELD dataset, the IEMOCAP dataset has a relatively small number of utterances and mainly records dyadic conversations.

From Table 4, we can first observe the poor performance of the single visual model. The single textual model works better than the single visual model. This phenomenon may be because the abstraction of visual sentiment makes it difficult for the CNN to find relatively good local optima. The FMF model can produce improved results over the single visual model but fails to improve the performance over the single textual model. This finding proves that the simple feature-level fusion method cannot effectively capture the correlation between multimodalities. On the other hand, the DSEF model improves the performance

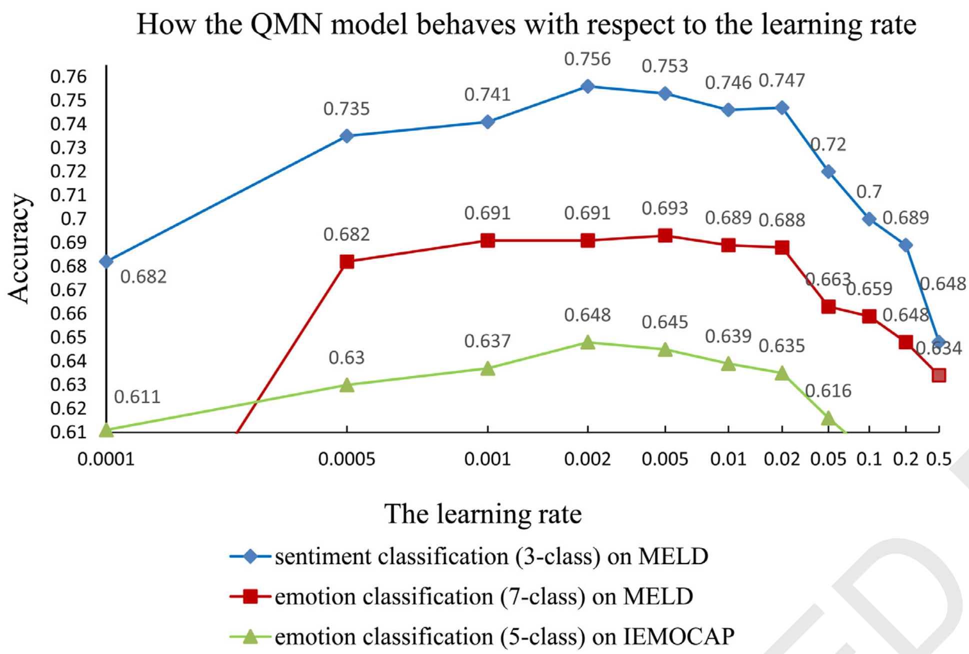

Fig. 6. How the QMN model behaves with respect to the learning rate.

in terms of both the weighted F1 score and the accuracy over the two single models. The D-S evidence theory can make good use of the prediction probabilities on text and image modalities to focus more on consistent information and eliminate contradictory information. Moreover, our defined mass function might be suitable for the IEMOCAP dataset. As one traditional deep learning-based multimodal method, the MDL model performs better than the FMF model, showing that learning a joint representation can improve sentiment classification performance.

In addition, the CRNN model outperforms the MDL model because the RNN can deal with sequences of conversational flows using its internal state (memory). However, this crude mechanism might not capture enough contextual information. Both the contextual and hierarchical contextual h-LSTM network models stably outperform all the other baselines because of their consideration of the contextual relations among the utterances. As a pioneering study in combining quantum representation with machine learning, the QMSA framework underperforms the two contextual h-LSTM network models but outperforms than the other baselines. Deep neural networks have stronger learning ability than SVM and RF models. The QMSA framework ignores the inter-utterance interactions. However, it still demonstrates the potential of using quantum theory as a formal framework for capturing lexical meaning.

Compared with the single textual and visual models, the accuracy results achieved by the textual and visual DM-CNN models increased by 4.43% and 3.94%, respectively. After being calculated, there are approximately 269 utterances misclassified by the single textual model that are accurately recognized by the textual DM-CNN model. There are approximately 210 visual utterances misclassified by the single visual model that are accurately recognized by the visual DM-CNN model. The textual and visual DM-CNN models outperform both the single textual and visual models, which shows the effectiveness of modeling term dependencies. In quantum theory, all the information contained in one system (which, in this paper, corresponds to each utterance) could be represented by the probability distribution of the measurement results and is embedded into the state space represented by the density matrix. Hence, the density matrix describes all the information and properties of the utterance. Density matrix-based CNN representation is an effective feature extraction approach that can be applied in text or image processing tasks. Through modeling the interaction between textual and visual predictions, the DM-QIMF model performs very well. The proposed quantum interference-inspired decision-level fusion method has taken the information-conflicting phenomenon that occurs in the process of multimodal information fusion into consideration.

Finally, aiming to establish an integrated theoretical system of quantum-like interaction modeling, our QMN model outperforms the hierarchical contextual h-LSTM network model by 3.7% in terms of the accuracy and 3.3% in terms of the F1 score. We think that this enhancement is caused by the fundamental differences between the QMN and contextual h-LSTM network models, which are reflected in three aspects: a) multimodal representation learning through a density matrix-based CNN; b) strong and weak interaction modeling; and c) decision fusion of multimodal sentiment labels.

5.5. Discussion of the learning rate

In this subsection, we search for the best performance from a parameter pool, which contains a learning rate in $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 2e^{-3}, 5e^{-3}, 1e^{-2}, 2e^{-2}, 5e^{-2}, 1e^{-1}, 2e^{-1}, 5e^{-1}\}$. We show how the QMN model behaves with respect to the learning rate on the MELD and IEMOCAP datasets. From Fig. 6, we notice that as the learning rate increases, the accuracy of our proposed QMN model increases in the first stage and then decreases on the three emotion recognition tasks. When we set the learning rate to $1e^{-4}$ and $5e^{-4}$, the QMN model does not perform well. This finding indicates that a smaller learning rate might lead the model to fall into a suboptimal solution. When the learning rate is set to $1e^{-1}$, $2e^{-1}$ and $5e^{-1}$, the performance of the QMN model falls sharply. An excessively large learning rate might lead to weight updates that will be too large, and gradient descent might increase rather than decrease the training error.

Finally, when we set the learning rate to $2e^{-3}$, the QMN model achieves the best performance on the 3-class and 5-class sentiment classification tasks. When the learning rate is set to $5e^{-3}$, the QMN model achieves the highest accuracy result on the 5-class classification task and outperforms the second highest result (which corresponds to a learning rate of $2e^{-3}$) by 0.29%. A well-configured learning rate helps the model approximate the function as closely as possible. Hence, taking the three comprehensive classification tasks into consideration, the learning rate is set to $2e^{-3}$ for the 3-class and 5-class classifications tasks and $5e^{-3}$ for the 7-class classification tasks.

5.6. Ablation study

In this subsection, we design a series of submodels for a comprehensive study on the impact of different components of the QMN model: (1) a DM-LSTM network, which does not model influences but only uses a density matrix-based CNN to extract textual and visual features, feeds them into two standard LSTM networks and fuses their local predictions using the standard linear combination method; (2) an influence-LSTM network, which uses standard CNNs to extract the textual and visual features, feeds them into two LSTM networks that have incorporated influences into the output gate and fuses their local predictions using the standard linear combination method; and (3) a QIMF-LSTM network, which uses standard CNNs to extract textual and visual features, feeds them into two standard LSTM networks and fuses their local predictions using the proposed quantum interference-inspired multimodal fusion approach.

From Table 5, we observe that the QMN model achieves the best performance among all the models. The results verify that modeling both the intra- and inter-utterance interactions makes a positive contribution to judging the sentiment polarity of an utterance. The DM-LSTM network model performs best among the three submodels, showing that the density matrix representation plays the most important role in improving performance. This importance is because the density matrix representation can more effectively encode the semantic dependencies and their probabilistic distribution information. However, we notice that the DM-LSTM network model might be sensitive to the large number of classes. The influence-LSTM network model attains the worst results among all submodels but still outperforms the CRNN, MDL and other baselines that ignore the interdependencies among utterances. This finding shows that modeling inter-utterance interactions benefits the sentiment classification performance. The QIMF-LSTM network model performs better than influence-LSTM but worse than DM-LSTM. Compared with the QMN model, it only uses the QIMF strategy to fuse the local decisions on texts and images that have been predicted by two LSTM networks, which means that the quantum interference-inspired decision fusion strategy is an effective fusion strategy, which is also rooted in a

Table 5
Ablated QMN for both MELD and IEMOCAP datasets.

Dataset	Model	Metric	
		F1	Accuracy
MELD	DM-LSTM	0.710	0.736
	Influence-LSTM	0.688	0.707
	QIMF-LSTM	0.699	0.711
	QMN	0.729	0.756
IEMOCAP	DM-LSTM	0.604	0.632
	Influence-LSTM	0.592	0.613
	QIMF-LSTM	0.597	0.625
	QMN	0.623	0.648

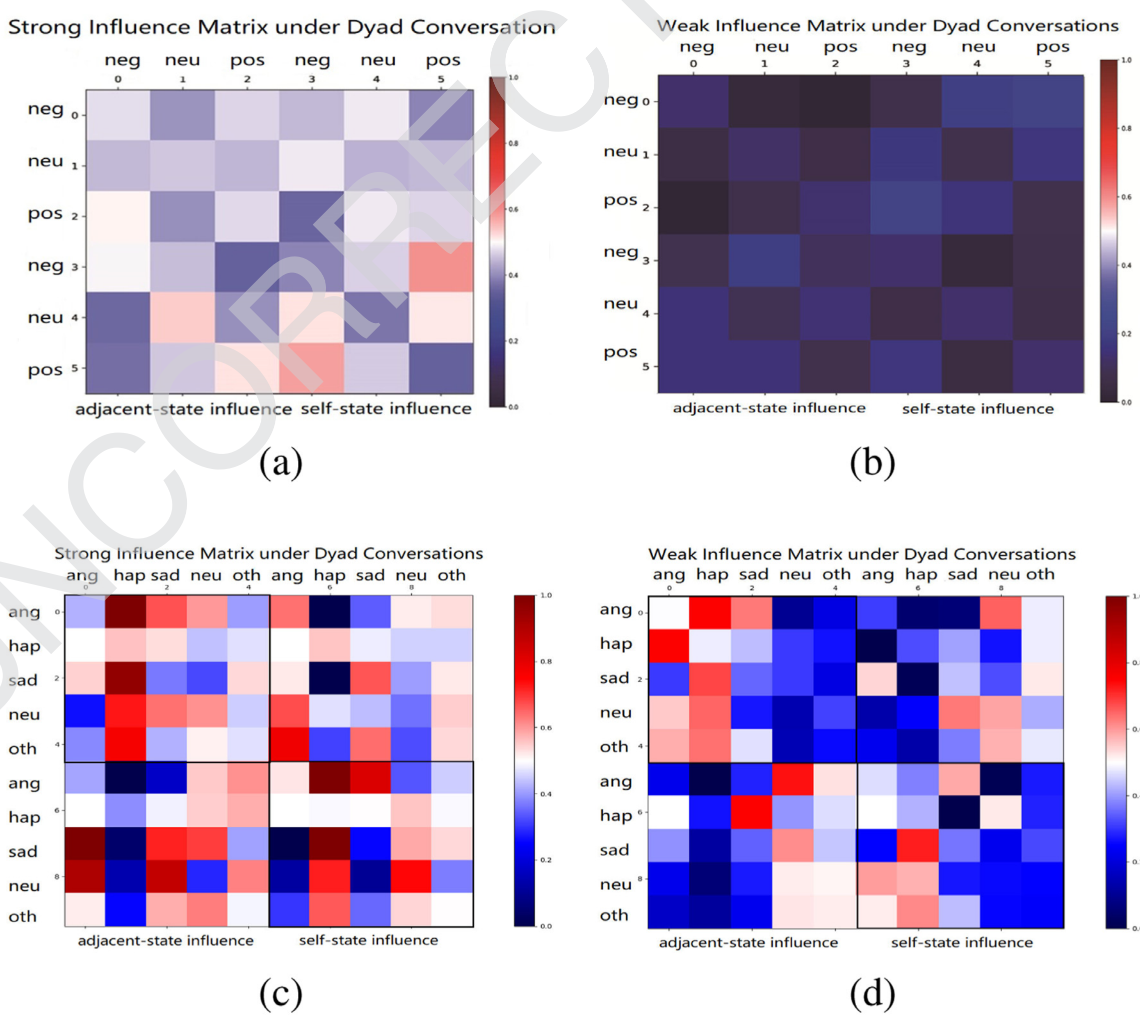

Fig. 7. (a) Strong influence matrix in the MELD dataset; (b) weak influence matrix in the MELD dataset; (c) strong influence matrix in the IEMOCAP dataset; (d) weak influence matrix in the IEMOCAP dataset. Different colors denote different influences.

well-founded mathematical derivation. Overall, the ablation study suggests that a) an effective semantic learning model could help the machine to better "understand" multimodal documents; b) influence matrices can effectively capture strong and weak dependency; and c) the QIMF strategy indeed incorporates some complementary decision information.

5.7. Visualization of the influence matrix and remarks

Fig. 7 demonstrates a way to visualize the influence matrices that allow us to observe strong and weak influences. Figs. 7(a) and 7(b) present two different types of influences (i.e., strong and weak, respectively) derived from the 3-class sentiment classification task on the MELD dataset; 0, 1 and 2 denote the *negative*, *neutral* and *positive* sentiments, respectively, of the first speaker, while 3, 4 and 5 denote the *negative*, *neutral* and *positive* sentiments of another speaker, respectively. Each image can be divided into 4 submatrices of size 3×3. The submatrices in the upper-left portion and lower-right portion represent the self-state influence. The submatrices in the upper-right portion and lower-left portion represent the adjacent-state influence.

Similarly, Figs. 7(c) and 7(d) present strong and weak influences derived from the task of 5-class emotion classification on the IEMOCAP dataset; 0, 1, 2, 3, 4 denote the *anger*, *happiness*, *sadness*, *neutral* and *others* sentiments, respectively, of the first speaker, while 5, 6, 7, 8 and 9 denote the same sentiments of another speaker. Each image can also be divided into four 5×5 submatrices. The submatrices positioned in the upper-left portion and lower-right portion represent the self-state influence. The submatrices positioned in the upper-right portion and lower-left portion represent the adjacent-state influence.

Overall, we see that the tones of Fig. 7(a) are pale white, mixing a hint of light red, while those of Fig. 7(b) are more in the black and blue zones. This finding indicates that the strong influence matrix does capture stronger influences, whose average values vary from 0.4 to 0.6, while the weak influence matrix does capture less strong (weaker) influences, whose average value is approximately 0.2. Similarly, Fig. 7(c) displays more red zones, while Fig. 7(d) contains more dark blue zones. Their average values are approximately 0.52 and 0.34, corresponding to strong and weak influences.

Specifically, for the strong influence matrix in the MELD, dataset we can see light-red zones positioned in the lower portion, which indicates that the latter speaker is greatly influenced by previous speakers and has a great influence on him or herself. For the weak influence matrix in the MELD dataset, we can see from the widely spread black zones that each speaker has a weak influence on others and him or herself. For the strong influence matrix in the IEMOCAP dataset, we can notice that more red zones are positioned in the upper portion. This finding indicates that the first speaker, who controls the rhythm of the conversation, has a great influence on him or herself and is moderately affected by another participant. For the weak influence matrix in the IEMOCAP dataset, we can observe an interesting phenomenon in which many blue zones are positioned in the tail in the top-left corner and the lower-right corner of the influence matrix, while the opposite is true for the top-right corner and the lower-left corner. This finding shows that the speaker who exhibits "neutral" and "others" emotions weakly affects him or herself. The speaker who exhibits the emotions of "happiness", "sadness" and "anger" is weakly affected by the other speaker.

5.8. Remarks on cos θ

The cos θ of the interference term comes from the phase of the product $\alpha\varphi_{text}(x)\cdot\beta\varphi_{img}(x)$, which can range from -1 to +1. In our work, we denote -1 as the most negative cognitive interference between the text and the image and denote +1 as the most positive cognitive interference. When $\cos\theta = 0$, we consider that there is no cognitive interference. In this subsection, we tune cos θ with different settings for an in-depth understanding of the impact of cos θ. Fig. 8 shows the impact of cos θ

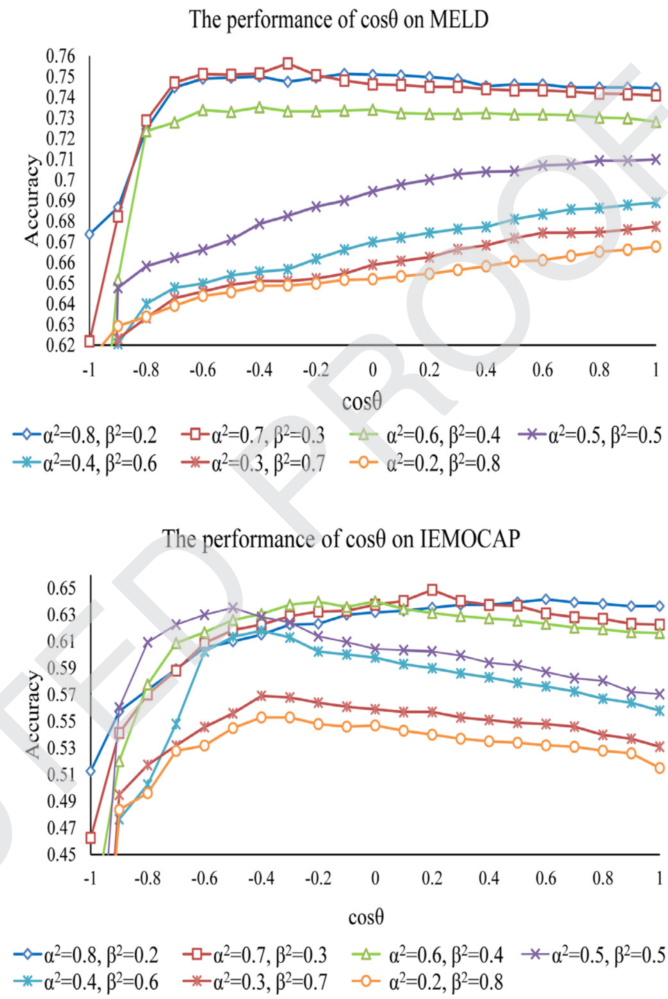

Fig. 8. The effect of cos θ on the MELD and IEMOCAP datasets.

on the MELD and IEMOCAP datasets. Note that in this paper, we set the single α and β values to deal with all the textual and visual predictions at once. Actually, we have also noticed that adjusting different α and β values for different multimodal utterances may reduce the number of false positives and further improve the performance. However, this strategy will increase the computational cost of the calculations. Considering the trade-off between effectiveness and computational burden, we only adjust the fixed α and β values in the current work.

We analyze how our QMN model behaves on the MELD and IEMOCAP datasets with respect to the parameter cos θ in light of different values of α and β. For the MELD dataset, the accuracy increases along with the increase in cos θ. Specifically, we can observe that the accuracy is the highest when $\alpha^2 = 0.7$ and $\beta^2 = 0.3$ on the MELD dataset. When $\alpha^2 = 0.2$ and $\beta^2 = 0.8$, the accuracy is the lowest. When $\alpha^2 = 0.6$ and $\beta^2 = 0.4$, the accuracy increases until $\cos\theta = -0.8$ and then remains almost unchanged. When $\alpha^2 = 0.5$ and $\beta^2 = 0.5$, $\alpha^2 = 0.4$ and $\beta^2 = 0.6$, $\alpha^2 = 0.3$ and $\beta^2 = 0.7$, and $\alpha^2 = 0.2$ and $\beta^2 = 0.8$, the accuracy increases until $\cos\theta = 1$. Furthermore, our QMN model achieves the best performance when $\cos\theta = -0.3$. This finding implies that there exists some weak negative interference between textual and visual sentiment recognition in the MELD dataset.

For the IEMOCAP dataset, when $\alpha^2 > \beta^2$, the accuracy increases along with the increase in cos θ. When $\alpha^2 \leq \beta^2$, the accuracy first increases and then decreases with increasing cos θ. A likely reason for this phenomenon is that all subjective videos in the IEMOCAP dataset only record two speakers who are in the same scenarios. Visual sentiment analysis is more difficult than textual sentiment analysis. If we pay more attention to images than texts, negative interference might improve the performance, while positive interference does not benefit the classification. However, if we pay more attention to the texts, the opposite is true. We can notice that when $\cos\theta = 0.2$, our QMN model achieves the best performance.

Our QMN model achieves the best performance on the MELD dataset when $\cos\theta = -0.3$, while it attains the best classification results on the IEMOCAP dataset when $\cos\theta = 0.2$. The videos in the MELD dataset are collected from a TV sitcom, in which actors/speakers wearing clothes of various styles and colors talk to each other in different scenarios. Videos in the IEMOCAP dataset only record two speakers wearing unchanging clothes in front of a single background. As a consequence, visual sentiment analysis on the MELD dataset is more difficult than it is on the IEMOCAP dataset, and we could also observe a similar phenomenon by comparing with their classification results from Table 3 and Table 4. Hence, the visual prediction results are usually in contrast to the textual prediction results on the MELD dataset, leading to weak negative interference. However, the visual prediction results usually coincide with the textual prediction results on the IEMOCAP dataset. Assigning $\cos\theta$ to a positive value might help improve the performance. Moreover, we can also see that the accuracy is highest when $\alpha^2 = 0.7$ and $\beta^2 = 0.3$. When $\alpha^2 = 0.2$ and $\beta^2 = 0.8$, the accuracy is the lowest. These two results indicate that analyzing the sentiment of a text input is probably more important in multimodal sentiment analysis than visual input.

6. Conclusions and future work

Conversational sentiment analysis is an important and challenging task. In this paper, we design a quantum-like multimodal network (QMN) framework, which leverages the mathematical formalism of quantum theory (QT) and a long short-term memory (LSTM) network, to model both intra- and inter-utterance interaction dynamics and recognize speakers' emotions. The main idea is to use a density matrix-based CNN, a quantum measurement-inspired strong-weak influence model and a quantum interference-inspired multimodal decision fusion approach. The experimental results on the MELD and IEMOCAP datasets demonstrate that our proposed QMN largely outperforms a wide range of baselines and state-of-the-art multimodal sentiment analysis algorithms, thus verifying the effectiveness of using quantum theory formalisms to model inter-utterance interaction, the fusion of multimodal contents and the fusion of local decisions (i.e., intra-utterance interactions).

Since the QMN model is largely dependent on the density matrix representation, how to take a further step towards accurately capturing the interactions among speakers and naturally incorporating them into an end-to-end framework will be left to our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yazhou Zhang: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Dawei Song:** Conceptualization, Validation, Writing - review & editing, Funding acquisition, Supervision. **Xiang Li:** Validation, Writing - review & editing, Data curation. **Peng Zhang:** Visualization, Project administration, Resources. **Panpan Wang:** Resources, Data curation. **Lu Rong:** Formal analysis. **Guangliang Yu:** Funding acquisition. **Bo Wang:** Funding acquisition.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (grant No. 2018YFC0831704), the Natural Science Foundation of China (grant No. U1636203, 61772363), the Major Project of Zhejiang Lab (grant No. 2019DH0ZX01), the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 721321, the National Natural Science Foundation of China (grant No. U1736103, U1504608, 61802352), the Foundation and Cutting-Edge Technologies Research Program of Henan Province (grant No. 192102210294). The Project of Science and Technology in Henan Province under Grant No. 202102210178, and the Industrial Science and Technology Research Project of Henan Province under Grants 202102210387.

Appendix A

Inference process of the strong-weak influence model

Given the model description and hyperparameter J, the likelihood function can be determined as:

$$\begin{aligned}
&\zeta\left(o_{1:T}^{1:C}, q_{1:T}^{1:C} \middle| E^{1:C}, F^{1:C}, R(1:J), r_{1:T}\right)\\
&= \prod_{e}^{C} P(o_1^e|q_1^e)P(q_1^e)\\
&\times \prod_{u=2}^{T} \{P(r_u)P_{e=1}(o_u^e|q_u^e)P(q_u^e|q_{u-1}^1, q_{u-1}^2, \dots, q_{u-1}^C)\\
&\times \prod_{e=2}^{C} P(o_u^e|q_u^e)P(q_u^e|q_u^1, q_u^2, \dots, q_u^{e-1}, q_{u-1}^e, q_{u-1}^{e+1}, \dots, q_{u-1}^C)\}
\end{aligned} \tag{A.1}$$

Depending on whether the training set contains the state sequence or not, the learning algorithm of HMM or its variants is divided into two categories of approaches: supervised learning and unsupervised learning. Since well-labeled training data are usually very expensive and time consuming to construct, unsupervised learning is the most commonly used method, such as the forward-backward and variational expectation maximization (EM) algorithms. The dynamic influence model adopts the EM approach to learn the parameters. However, in this paper, we used two well-labeled conversational datasets, which contain both the observation and the state sequence. Hence, we choose to use a supervised approach to learn the system parameters. Supervised learning estimates the transition/emission probabilities from known samples via the counting frequencies. Assume that there are two speakers A and B in a conversation; i.e., entity $C = 2$, and the first speaker of each turn u is A, the second speaker is B. The inference process is as follows.

$$E_{s_i,s_j}^{e}|_{e\in\{A,B\}} = \frac{\sum_u Count\left(q_u^e = s_i, q_{u+1}^e = s_j\right)}{\sum_u \sum_s Count\left(q_u^e = s_i, q_{u+1}^e = s\right)} \tag{A.2}$$

$$F_{s_i,s_j}^{B} = \frac{\sum_u Count\left(q_u^B = s_i, q_{u+1}^A = s_j\right)}{\sum_u \sum_s Count\left(q_u^B = s_i, q_{u+1}^A = s\right)} \tag{A.3}$$

$$F_{s_i,s_j}^{A} = \frac{\sum_u Count\left(q_u^A = s_i, q_u^B = s_j\right)}{\sum_u \sum_s Count\left(q_u^A = s_i, q_u^B = s\right)} \tag{A.4}$$

$$R_{e_1,e_2}^{j} = \begin{cases} \frac{F_{si,s_j}^{e_2}}{E_{s_i,sj}^{e_1}+F_{si,s_j}^{e_2}} & e_1 \neq e_2, r_t = j \\ \frac{E_{s_i,sj}^{e_1}}{E_{s_i,sj}^{e_1}+F_{s_i,s_j}^{e'}} & (8) e_1 = e_2, e' = C - e_1, r_t = j \end{cases} \tag{A.5}$$

and the emission probability is

$$b_{s_j}(o_k) = \frac{\sum_u Count\left(q_u^e = s_j, o_u^e = o_k\right)}{\sum_u \sum_o Count(q_u^e = s_j, o_u^e = o)} \tag{A.6}$$

References

[1] S. Kumar, M. Yadava, P.P. Roy, Fusion of eeg response and sentiment analysis of products review to predict customer satisfaction, Inf. Fusion 52 (2019) 41–52.

[2] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[3] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.
[4] X. Liu, Y. Xu, F. Herrera, Consensus model for large-scale group decision making based on fuzzy preference relation with self-confidence: detecting and managing overconfidence behaviors, Inf. Fusion 52 (2019) 245–256.
[5] M. Dragoni, S. Poria, E. Cambria, Ontosenticnet: a commonsense ontology for sentiment analysis, IEEE Intell. Syst. 33 (3) (2018) 77–85.
[6] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image Vis. Comput. 65 (2017) 3–14.
[7] Y. Qian, Y. Zhang, X. Ma, H. Yu, L. Peng, Ears: emotion-aware recommender system based on hybrid information fusion, Inf. Fus. 46 (2019) 141–146.
[8] J.A. Balazs, J.D. Velásquez, Opinion mining and information fusion: a survey, Inf. Fusion 27 (2016) 95–110.
[9] I. Chaturvedi, E. Cambria, R.E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: survey and challenges, Inf. Fusion 44 (2018) 65–77.
[10] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.
[11] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114.
[12] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: addressing key issues and setting up the baselines, IEEE Intell. Syst. 33 (6) (2018) 17–25.
[13] P. Zhang, Z. Su, L. Zhang, B. Wang, D. Song, A quantum many body wave function inspired language modeling approach, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, 2018, pp. 1303–1312.
[14] Y. Wang, C. von der Weth, Y. Zhang, K.H. Low, V.K. Singh, M. Kankanhalli, Concept based hybrid fusion of multimodal event signals, in: 2016 IEEE International Symposium on Multimedia (ISM), IEEE, 2016, pp. 14–19.
[15] V.P. Rosas, R. Mihalcea, L.-P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.
[16] Q. You, J. Luo, H. Jin, J. Yang, Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia, in: Proceedings of the Ninth ACM international conference on Web search and data mining, ACM, 2016, pp. 13–22.
[17] Q. Yang, Y. Rao, H. Xie, J. Wang, F.L. Wang, W.H. Chan, E.C. Cambria, Segment-level joint topic-sentiment model for online review analysis, IEEE Intell Syst 34 (1) (2019) 43–50.
[18] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intell. Syst. 32 (6) (2017) 74–80.
[19] C. Welch, V. Pérez-Rosas, J.K. Kummerfeld, R. Mihalcea, Learning from personal longitudinal dialog data, IEEE Intell Syst 34 (4) (2019) 16–23.
[20] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 873–883.
[21] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, 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, 1, 2018, pp. 2122–2132.
[22] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1, 2019, pp. 527–536.
[23] B. Ojamaa, P.K. Jokinen, K. Muischenk, Sentiment analysis on conversational texts, in: Proceedings of the 20th Nordic Conference of Computational Linguistics, Linköping University Electronic Press, 2015, pp. 233–237. 109.
[24] J. Bhaskar, K. Sruthi, P. Nedungadi, Hybrid approach for emotion classification of audio conversation based on text and speech mining, Procedia Comput. Sci. 46 (2015) 635–643.
[25] A. Sordoni, J.-Y. Nie, Y. Bengio, 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, ACM, 2013, pp. 653–662.
[26] P. Wang, T. Wang, Y. Hou, D. Song, Modeling relevance judgement inspired by quantum weak measurement, in: European Conference on Information Retrieval, Springer, 2018, pp. 424–436.
[27] Y. Zhang, D. Song, X. Li, P. Zhang, Unsupervised sentiment analysis of twitter posts using density matrix representation, in: European Conference on Information Retrieval, Springer, 2018, pp. 316–329.
[28] Q. Li, J. Li, P. Zhang, D. Song, Modeling multi-query retrieval tasks using density matrix transformation, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2015, pp. 871–874.
[29] P. Zhang, J. Niu, Z. Su, B. Wang, L. Ma, D. Song, End-to-end quantum-like language models with application to question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5666–5673.
[30] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 5436–5442, doi:10.24963/ijcai.2019/755.
[31] R. Ji, D. Cao, D. Lin, Cross-modality sentiment analysis for social multimedia, in: Multimedia Big Data (BigMM), 2015 IEEE International Conference on, IEEE, 2015, pp. 28–31.
[32] H. Abburi, E.S.A. Akkireddy, S. Gangashetti, R. Mamidi, Multimodal sentiment analysis of telugu songs, in: Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), 2016, pp. 48–52.
[33] Y. Yoshitomi, S.-I. Kim, T. Kawano, T. Kilazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in: Robot and Human Interactive Communication, RO-MAN 2000, IEEE, 2000, pp. 178–183.
[34] M. Pantic, N. Sebe, J.F. Cohn, T. Huang, Affective multimodal human-computer interaction, in: Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 669–676.
[35] A. Mehrabian, Communication without words, Communication Theory 2 (2008) 193–200.
[36] N. Sebe, I. Cohen, T. Gevers, T.S. Huang, Emotion recognition based on joint visual and audio cues, in: 18th International Conference on Pattern Recognition, ICPR 2006, 1, IEEE, 2006, pp. 1136–1139.
[37] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, 2011, pp. 169–176.
[38] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 1, 2013, pp. 973–982.
[39] D. Gkoumas, D. Sogn, Exploiting quantum-like interference in decision fusion for ranking multimodal documents, ArXiv abs/1811.11422 (2018) 1–12.
[40] Q. Li, Multimodal data fusion with quantum inspiration, in: Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, in: SIGIR'19, 2019, p. 1451.
[41] Q. Li, M. Melucci, Quantum-inspired multimodal representation, in: 10th Italian Information Retrieval Workshop, 2019, pp. 1–2.
[42] Q. Li, B. Wang, M. Melucci, 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, NAACL-HLT 2019, 2019, pp. 4139–4148.
[43] E. Cambria, S. Poria, A. Hussain, Speaker-independent Multimodal Sentiment Analysis for Big Data, in: Multimodal Analytics for Next-Generation Big Data Technologies and Applications, Springer, 2019, pp. 13–43.
[44] A. Kumar, G. Garg, Sentiment analysis of multimodal twitter data, Multimed. Tools Appl (2019) 1–17.
[45] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks., in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 381–388.
[46] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 163–171.
[47] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: Data Mining (ICDM), 2017 IEEE International Conference on, IEEE, 2017, pp. 1033–1038.
[48] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl. Based Syst. 167 (2019) 26–37.
[49] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intell. Syst. 34 (3) (2019) 38–43, doi:10.1109/MIS.2019.2904691.
[50] J. Yu, J. Jiang, Adapting bert for target-oriented multimodal sentiment classification, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 2019, pp. 5408–5414, doi:10.24963/ijcai.2019/751.
[51] S. Verma, C. Wang, L. Zhu, W. Liu, Deepcu: integrating both common and unique latent information for multimodal sentiment analysis, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 3627–3634.
[52] N. Xu, W. Mao, C. Guandan, Multi-interactive memory network for aspect based multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 371–379, doi:10.1609/aaai.v33i01.3301371.
[53] M. Pagé Fortin, B. Chaib-draa, Multimodal multitask emotion recognition using images, texts and tags, in: Proceedings of the ACM Workshop on Crossmodal Learning and Application, ACM, 2019, pp. 3–10.
[54] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270.
[55] M. Huddar, S. Sannakki, V. Rajpurohit, A survey of computational approaches and challenges in multimodal sentiment analysis, Int. J. Comput. Sci. Eng. 7 (2019) 876–883, doi:10.26438/ijcse/v7i1.876883.
[56] S.H. Dumpala, I. Sheikh, R. Chakraborty, S.K. Kopparapu, Audio-visual fusion for sentiment classification using cross-modal autoencoder, in: 32nd Conference on Neural Information Processing Systems (NIPS 2018), NIPS, 2018, pp. 1–4.
[57] E. Russell, Real-time topic and sentiment analysis in human-robot conversation, Master's Theses 1 (2015) 338–729.
[58] D. Mahata, J. Friedrichs, R.R. Shah, J. Jiang, Detecting personal intake of medicine from twitter, IEEE Intell. Syst. 33 (4) (2018) 87–95.
[59] S. Maghilnan, M.R. Kumar, Sentiment analysis on speaker specific speech data, in: Intelligent Computing and Control (I2C2), 2017 International Conference on, IEEE, 2017, pp. 1–5.
[60] E. Hoque, G. Carenini, Convis: A visual text analytic system for exploring blog conversations, in: Computer Graphics Forum, 33, Wiley Online Library, 2014, pp. 221–230.

[61] M. Mazzocut, I. Truccolo, M. Antonini, F. Rinaldi, P. Omero, E. Ferrarin, P. De Paoli, C. Tasso, Web conversations about complementary and alternative medicines and cancer: content and sentiment analysis, J. Med. Internet Res. 18 (6) (2016) 221–230.
[62] C. Bothe, S. Magg, C. Weber, S. Wermter, Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 477–485.
[63] E. Huijzer, Identifying effective affective email responses, Master Thesis Business Analytics (2017) 1–75.
[64] A. Aznar, H.R. Tenenbaum, Gender comparisons in mother-child emotion talk: ameta-analysis, Sex Roles (2019) 1–8.
[65] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 6818–6825.
[66] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 5415–5421.
[67] P. Zhong, D. Wang, C. Miao, Knowledge-enriched transformer for emotion detection in textual conversations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 165–177.
[68] Z. Rebiai, S. Andersen, A. Debrenne, V. Lafargue, Scia at semeval-2019 task 3: Sentiment analysis in textual conversations using deep learning, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 297–301.
[69] J.v. Neumann, Mathematische grundlagen der quantenmechanik, 38, Springer-Verlag, 2013.
[70] N. Bourbaki, Elements of Mathematics: General Topology, 3, Hermann, 1966.
[71] P. Busch, Quantum states and generalized observables: a simple proof of gleason??s theorem, Phys. Rev. Lett. 91 (12) (2003) 120–403.
[72] L. Masanes, M.P. Müller, A derivation of quantum theory from physical requirements, New J Phys 13 (6) (2011) 1–63.
[73] J. Von Neumann, Mathematical foundations of quantum mechanics: New edition, Princeton university press, 2018.
[74] M. Sands, R.P. Feynman, R. Leighton, The feynman lectures on physics: Mainly electromagnetism and matter, 2017.
[75] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[76] W. Pan, W. Dong, M. Cebrian, T. Kim, J.H. Fowler, A.S. Pentland, Modeling dynamical influence in human interaction: using data to make better inferences about influence within social systems, IEEE Signal Process Mag. 29 (2) (2012) 77–86.
[77] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[78] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 325–335.
[79] S. Bird, E. Klein, E. Loper, Natural language processing with python: Analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009.
[80] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 13–28.
[81] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: OSDI, 16, 2016, pp. 265–283.
[82] D. Pan, P. Zhang, J. Li, D. Song, J.-R. Wen, Y. Hou, B. Hu, Y. Jia, A. De Roeck, Using dempster-shaferâs evidence theory for query expansion based on freebase knowledge, in: Asia Information Retrieval Symposium, Springer, 2013, pp. 121–132.
[83] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.
[84] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, B. Hu, Emotion recognition from multi-channel eeg data through convolutional recurrent neural network, in: Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, IEEE, 2016, pp. 352–359.