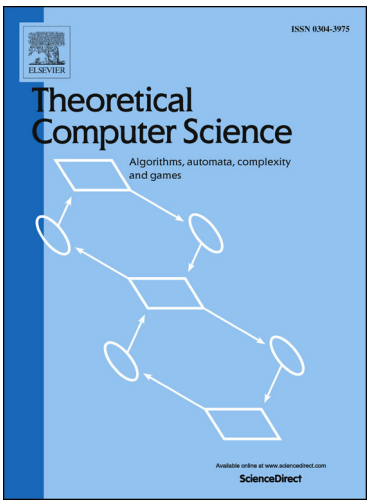
A Quantum-Inspired Multimodal Sentiment Analysis Framework

Accepted Manuscript



A Quantum-Inspired Multimodal Sentiment Analysis Framework

Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li et al.

PII: S0304-3975(18)30263-9
DOI: https://doi.org/10.1016/j.tcs.2018.04.029
Reference: TCS 11559

To appear in: *Theoretical Computer Science*

Received date: 12 June 2017
Revised date: 29 March 2018
Accepted date: 13 April 2018

Please cite this article in press as: Y. Zhang et al., A Quantum-Inspired Multimodal Sentiment Analysis Framework, *Theoret. Comput. Sci.* (2018), https://doi.org/10.1016/j.tcs.2018.04.029

A Quantum-inspired Multimodal Sentiment Analysis Framework

Yazhou Zhang[a], Dawei Song[b,c,*], Peng Zhang[a,*], Panpan Wang[a], Jingfei Li[a], Xiang Li[a], Benyou Wang[a]

[a]*Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, 300350, No.135 Yaguan Road, Jinnan District, Tianjin, 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*

Abstract

Multimodal sentiment analysis aims to capture diversified sentiment information implied in data that are of different modalities (e.g., an image that is associated with a textual description or a set of textual labels). The key challenge is rooted on the "semantic gap" between different low-level content features and high-level semantic information. Existing approaches generally utilize a combination of multimodal features in a somehow heuristic way. However, how to employ and combine multiple information from different sources effectively is still an important yet largely unsolved problem. To address the problem, in this paper, we propose a Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework. The framework consists of a Quantum-inspired Multimodal Representation (QMR) model (which aims to fill the "semantic gap" and model the correlations between different modalities via density matrix), and a Multimodal decision Fusion strategy inspired by Quantum Interference (QIMF) in the double-slit experiment (in which the sentiment label is analogous to a photon, and the data modalities are analogous to slits). Extensive experiments are conducted on two large scale datasets, which are collected from the Getty Images and Flickr photo sharing platform. The experimental results show that our approach significantly outperforms a wide range of baselines and state-of-the-art methods.

Keywords: Multimodal sentiment analysis, Quantum theory, Decision fusion, Information fusion

1. Introduction

With the rapid development of WWW and social networking services, more and more people express their opinions and sentiments on social media platforms (e.g., Facebook, Twitter, etc.) by publishing user generated content and posting comments. Instead of using textual content only, people nowadays are willing to share opinions through multimodal content (e.g., text+image, text+video, audio+image), which are expected to express personal feelings more accurately and intuitively than the use of mono-modal content, as illustrated in Figure 1. The importance of analyzing the sentiment polarity of online multimodal content has been recognized in a wide range

*Corresponding authors: Dawei Song (dawei.song2010@gmail.com) and Peng Zhang (pzhang@tju.edu.cn)

of application domains, e.g., to help manufacturers improve their products and to help governments understand public opinions. Therefore, multimodal sentiment analysis is of a theoretical and practical significance, and has attracted an increasing attention from both academia and industry [1][2][3][4]. In this paper, we focus on identifying the overall sentiment of users implied in images and texts published on social media platforms.

(a) Sad clown in the rain (b) Sad fly (c) The dog and the music

Figure 1: Examples of Flickr multimodal documents. (a) Both text and image carry a negative sentiment; (b) it is difficult to identify sentiment from the image, but the text carries a negative sentiment; (c) conversely, it is easier to identify negative sentiment from the image instead of the text.

There has been a growing literature in multimodal sentiment analysis for social media, such as YouTube, Twitter, Microblog and Facebook. For instance, Baecchi et al. [5] adopted deep neural networks to analyze the sentiment of Twitter documents. Ji et al. [6] proposed to use hypergraph to model the correlations among different modalities on Sina Microblog data. Mihalcea et al. [7] combined textual, visual and acoustic features to identify the sentiment expressed in utterance-level visual data streams from YouTube. There is also a large body of work on other social media platforms, e.g., Getty Images [8], Newscast [9], and Telugu Songs [10].

Nowadays, most multimodal sentiment analysis methods focus on extracting effective features or training a robust classifier. Several studies adopt simple decision fusion strategies (such as a linear combination or voting strategy) to model the correlation among multi-modalities at the decision level [11][12]. Despite of the remarkable progress that has been made, how to effectively employ and combine multiple modalities of information from different sources is still an important yet largely unsolved problem.

Moreover, visual sentiment analysis is also a challenging task since it involves a higher level of abstraction and subjectivity than textual sentiment. The well-known "semantic gap" needs to be filled. For an image, a basic representation is its pixel matrix. However, the pixel matrix does not carry higher-level semantic information. Therefore, capturing visual semantic information is a basis for effective multimodal sentiment analysis. Furthermore, multimodal sentiment analysis involves a complex decision process, in which different modalities often intertwine together to carry a common sentiment polarity. The sentiment information of different modalities will influence the final decision simultaneously. For example, the sentiment polarity of multimodal document is affected by the sentiment polarities of the text and its accompanying image.

To tackle these problems, we consider multimodal sentiment analysis within a more general mathatical framework. In the field of probability theory, both classical probability theory [13] and quantum probability theory [14] have been investigated. Classical probability, with its axiomatic foundation derived from the Kolmogorov's theory [15], has developed over centuries based on classic physics. Quantum probability theory, which is axiomatized by Von Neumann

based on quantum physics [16], has recently been shown to provide a more general framework for modelling a wide range of natural language understanding, information retrieval, user interaction and decision making problems [17][18][19][20]. For instance, classical probability theory obeys the commutative law, which means the order of different events does not matter. Obviously, this axiom is not sufficient to explain the above examples. Quantum probability theory [21][22], on the other hand, does not necessarily obey the commutative law. It thus provides a fresh conceptual framework for modeling multimodal sentiment analysis.

In this paper, we explore the use of Quantum Theory (QT) to model the multimodal sentiment analysis task, due to the following reasons: (a) as a theoretical framework that unifies probabilistic, logic and geometric formalisms, QT allows to consider the uncertainty in the process of decision making (such as combining decision information) [23]. (b) QT has been successfully applied to model various user-oriented aspects in ad-hoc information retrieval [24][25][26] and session search [27]. The intrinsic connections between QT and multimodal sentiment analysis as discussed above indicate that the insights and formalisms of quantum mechanics can be adopted to model multimodal sentiment analysis from a novel perspective [20][28][29].

Specifically, we propose a Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework. Our framework consists of two parts, including a representation learning model and a multimodal decision fusion strategy. These two parts can either integrate with each other as a whole for multimodal sentiment analysis or deal with the semantic gap and the decision fusion problems separately.

In the first part, we propose a Quantum-inspired Multimodal Representation (QMR) model, which represents the multimodal content as density matrices. For images, individual pixels are meaningless for human understanding of an image unless they construct abstract visual semantics. In the proposed QMR model, the pixels of an image are firstly used to construct visual words. Then these visual words are mathematically modeled as projectors onto a vector space, which can be seen as a process of higher-level abstraction. Finally, these projectors are encapsulated in a density matrix that describes a probability distribution of visual words of the image. For text, all words are modeled as projectors and are encapsulated in a density matrix in the similar way. Compared with the traditional vector-based representation model, the QMR model encodes more semantic information and naturally captures the inter-modal correlations.

In the second part, we investigate the information conflicting phenomenon that takes place in the process of multimodal information fusion. The ultimate goal of sentiment analysis is to enable the machine to correctly identify the sentiment polarity of information. As shown in Figure 1, Identifying final sentiment polarity produces information conflicting effect when it combines two sentiment decision information (which are image and text in our study). Therefore, this effect can be elaborated as an analogy to motivate our work. To address this challenge, drawing upon the double-slit experiment in quantum physics that demonstrate the quantum interference phenomenon, we propose a Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy. It is important to note that, we aim at developing a novel multimodal sentiment analysis model with the inspiration of QT, instead of explaining or modeling the state of mind of humans.

The main contributions of this paper can be summarized as follows:

- To our best knowledge, we are the first to apply Quantum Theory (QT) to sentiment analysis.
- We propose a Quantum-inspired Multimodal Representation (QMR) model to extract the semantic information of individual modalities, which are encapsulated in density matrices.

- We elaborate an analogy between multimodal sentiment analysis with a well-known double-slit experiment, and propose a Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy.
- We propose an integrated Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework, which contains the QMR model to fill the semantic gap and the QIMF strategy to fuse different decision results.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related work. Section 3 presents the proposed Quantum-inspired Multimodal Representation (QMR) model. In Section 4, we elaborate an analogy in multimodal sentiment analysis with the double-slit interference experiment, and describe the proposed Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy. In Section 5, we present our Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework that integrates the QMR model and the QIMF strategy. In Section 6, we report the empirical experiments. Section 7 concludes the paper and points out a number of future research directions.

2. Related Work

Generally speaking, there exist two categories of approaches in the current lierature of sentiment analysis (SA): lexicon-based (knowledge-based) approaches and machine learning based (statistical) approaches.

2.1. Lexicon-based Sentiment Analysis

The lexicon-based approaches infer the overall sentiment polarity of a piece of text based on the polarity of the words that compose it. These approaches depend on the sentiment dictionary and sentiment rules, which do not require storing a large data corpus and training algorithms. Early representatives in this category are Hatzivassiloglou [30] and Turney [31]. Later researchers have focused on using adjectives, adverbs or nouns as sentiment indicators of the text [32, 33]. Further research has involved building good dictionaries and judging the sentiment polarity of the text through the dictionaries. Some well-known dictionaries include SentiWordNet [34], MPQA [35] and GI [36].

Recent studies have been extended to sentiment analysis of online social media data. Musto et al. [37] proposed a lexicon-based approach for sentiment classification of Twitter posts. Moreno-Ortiz et al. [38] performed an evaluation using Sentitext, a lexicon-based SA tool for Spanish Twitter. Trinh et al. [39] built a Vietnamese emotional dictionary (VED) for sentiment analysis with Facebook data. Cui et al. [40] constructed a Weibo lexicon and used a propagation algorithm to automatically assign sentiment polarity scores to Chinese microblog messages. Saif et al. [41] proposed a semantic sentiment representation of words called SentiCircle, and performed entity- and tweet-level level sentiment analysis on Twitter data.

More recently, there are studies that combine natural language processing (NLP) and semantic web approaches for sentiment analysis [42, 43, 44, 45]. Semantic web based sentiment analysis can take advantage from linked data, ontologies, controlled vocabularies to deal with the domain-dependent problem. For example, Recupero et al. [44] developed a semantic SA system that is able to recognize the holder of an opinion to detect the sentiment of a sentence.

As the Lexicon-based methods largely depend on dictionaries, they are mainly focused on text sentiment analysis and difficult to be extended to other modalities. Moreover, the classification accuracy is generally lower than machine learning approaches, which are described next.

2.2. *Statistical Approaches to Sentiment Analysis*

The statistical approaches make use of machine learning methods, such as random forest, support vector machines, and neural networks. They involve building classifiers from labeled data, essentially a supervised classification task. Pang et al. [46] employed three machine learning methods to classify overall sentiment of documents. Pak [47] collected a Twitter corpus and build a sentiment classifier to determine sentiments of documents. There were also attempts in combining machine learning and lexicon based approaches to analyze sentiment, which achieved an improved accuracy [48, 49].

The statistical approaches have been applied to other modalities. The work in [50] used machine learning algorithms to predict the sentiment of images based on SIFT features. Asghar et al. [51] presented a brief survey on analysing sentiment of YouTube users, showing that there is still a long way to go to solve this problem. Recently, multimodal sentiment analysis has been emerging [52, 53, 54]. Morency and Mihalcea [55] integrated visual, audio and textual features to address the task of tri-modal sentiment analysis for the first time. They also conducted experiments on Spanish videos and utterance-level visual datasets using the similar idea [3]. Maynard et al. [56] suggested considering contextual information to help resolve ambiguity in multimodal sentiment analysis. Poria et al. [1] used both feature- and decision-level fusion methods to merge audio, visual and textual clues for YouTube. You et al. [8] proposed a cross-modality consistent regression (CCR) model to analyze Getty Images and Twitter multimedia content. Similar approaches have also been developed for the analysis of Sina microblog data [57, 58, 6].

The statistical approaches for sentiment analysis have benefited from the popularity and fast development of machine learning methods. In general, they can achieve a better performance than the lexicon-based approaches. However, they rely heavily on the labeled datasets and introduce higher computational-complexity.

To sum up, the afore-described two categories of approaches have made a good progress in sentiment analysis and motivated our work. However, the existing approaches mostly focus on on extracting effective features and constructing robust classifiers or studying refined sentimental rules. They lack a principled theoretical framework to fill the semantic gap and have rarely considered the multi-source information fusion problem in multimodal sentiment analysis. In this paper, we propose a novel quantum-inspired framework to address the above two challenges.

3. A Quantum-inspired Multimodal Representation Model for Representation Learning

In multimodal sentiment analysis, textual and visual words can be seen as events, and the texts and the images can be seen as systems (probability distributions over the events). In quantum probability theory, events are defined as projectors, systems are represented by density matrices on the probability space. This motivates us to propose a Quantum-inspired Multimodal Representation (QMR) model via density matrix.

3.1. Preliminaries of Quantum Theory

In QT, the quantum probability space is naturally encapsulated in an infinite Hilbert space, noted as $\mathbb{H}^n$. With the 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, $|\varphi\rangle = \sum_{i=1}^{n} a_i|e_i\rangle$, so does the wave function. Given two state vectors $|\varphi_1\rangle$ and $|\varphi_2\rangle$, the inner product between them is represented as $\langle\varphi_1|\varphi_2\rangle$. Similarly, the Hilbert space representation of the wavefunction is recovered from the inner product $\varphi(x) = \langle x|\varphi\rangle$.

In QT, assuming $|u\rangle$ is a unit vector, the projector Π on the direction u is writen as $|u\rangle\langle u|$. $|u\rangle\langle u|$ can also represent a density matrix of pure state. A real density matrix ρ is symmetric, $\rho = \rho^T$, positive semi-definite, $\rho \geq 0$, and of trace 1, i.e., $tr(\rho) = 1$. 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 $\{|e_i\rangle\}$, $\sum_{i=1}^{n} \mu(|e_i\rangle\langle e_i|) = 1$. The Gleason's Theorem has proven the existence of a mapping function $\mu(|u\rangle\langle u|) = tr(\rho|u\rangle\langle u|)$ for any vector $|u\rangle$.

The wave function and the density matrix to quantum theory are formally equivalent, each of which has its advantages in different applications. The wave function is good at describing the quantum mechanics of a particle through a wave-like description [59, 60]. The density matrix can more intuitively display the data distribution. In this paper, we employ the density matrix to extract the semantic information of multimodal data, and choose the wave function to formalize a decision fusion strategy.

3.2. the Quantum-inspired Multimodal Representation (QMR) Model

We propose a unified Quantum-inspired Multimodal Representation (QMR) model to represent the text and the image through density matrices. Aiming at an effective representation learning model, we base our computational framework on the Quantum Language Model (QLM) [61]. QLM is a novel application of quantum probability to information retrieval (IR), and achieves significant improvements over the classical probabilistic language model. In QLM, both single terms and compound term dependencies are modeled as projectors in a vector space. Documents and queries are represented as a sequence of projectors, encapsulated in density matrices. Although QLM is an effective text IR model, it is not suitable for multimodal sentiment analysis, especially for visual sentiment analysis. Moreover, the estimation of QLM may not always ensure a good convergence.

Different from classical probability theory, the events in QT are defined as subspaces, which are represented by any orthogonal projectors. All textual and visual words can be seen as events. Therefore, all single textual and visual words in a multimodal document are modeled as projectors Π. The projectors Π are used to estimate density matrices ρ of the corresponding multimodal document, which are probability distributions, corresponding to the probabilities of all events. Theoretically, compared with vector-based representation, density matrices can better encode the semantic dependencies and their probabilistic distribution information.

Specifically, for text, suppose $|w_i\rangle$ is a normalized word vector. The projector Π_i for a single word w_i is formulated in Eq. (1). One-hot representation of words over other words is known to suffer from the curse of dimensionality and difficulty in representing ambiguous words. We use word embeddings instead of one-hot representation to construct projectors in semantic space. In this paper, we employ the Glove tool [62] to find each word's embedding.

$$\Pi_i = |w_i\rangle\langle w_i| \tag{1}$$

For an image, we consider it as a document of visual words, in which each visual word is equivalent to a word in document. Therefore, we use these words $|s_i\rangle$ to represent projectors. This process is as described in the following procedure: (a) extracting SIFT features from all images in the training set; (b) clustering these extracted SIFT features to get k cluster centers through a k-means algorithm. Each cluster center is a visual word, and all k visual words form a visual dictionary; (c) using these visual words $|s_i\rangle$ to construct projectors $\Pi_i = |s_i\rangle\langle s_i|$ using Equation 1.

After defining projectors for each textual word and each visual word, we can represent a document with a sequence of projectors temporarily, $\mathcal{P}_U = \{\Pi_1, \Pi_2, ..., \Pi_n\}$, where n is the number of terms in the document. Then we use the Maximum Likelihood Estimation (MLE) to train density matrices ρ of documents and images as in the QLM. The likelihood function $\zeta(\rho)$ is the probability of getting the observed data given the density matrix:

$$\zeta(\rho) \propto \prod_i tr(\Pi_i \rho) \tag{2}$$

Since the log function is monotonic, the objective function $F(\rho)$ can be formulated as:

$$\begin{aligned} &F(\rho) \equiv \max_{\rho} \sum_i \log\left(tr(\Pi_i \rho)\right), \\ &subject\ to\ tr(\rho) = 1, \\ &\rho \geq 0 \end{aligned} \tag{3}$$

In the original QLM approach, an algorithm called RρR is used to estimate the maximum likelihood value. However, there is no theoretical guarantee of convergence, regardless the dataset and the initial value [63]. This algorithm may also suffer overshooting problem. To solve these problems, we employ a globally convergent algorithm [64], which extends the RρR algorithm. The ascent direction of likelihood is determined by two ascent directions controlled by the step size t. It is able to find a value which ensures a sufficient improvement in the likelihood function. It has been shown that an inexact line search method to determine t is enough for finding a value to guarantee the global convergence.

Specifically, based on the gradient of the objective function $F(\rho)$, this algorithm defines that $\nabla F(\rho) = \sum_i \frac{f_i}{tr(\Pi_i \rho)} \Pi_i$, where f_i is the term frequency. It also determines a definition that a direction D^k is an ascent direction at the kth iteration if $tr\left(\nabla F\left(\rho^k\right) D^k\right) > 0$, and this definition ensures that the function value increases.

The search direction D^k is a combination of the direction $\bar{D}^k$ and $\tilde{D}^k$, where $\bar{D}^k$ and $\tilde{D}^k$ are also ascent directions for any ρ^k. Using the Armijo condition and a backtracking procedure, the search direction D^k at the kth iteration is given by:

$$D^k = \frac{2}{q(t_k)} \bar{D}^k + \frac{t_k tr\left(\nabla F\left(\rho^k\right) \rho^k \nabla F\left(\rho^k\right)\right)}{q(t_k)} \tilde{D}^k \tag{4}$$

where $q(t_k)$, $\bar{D}^k$, $\tilde{D}^k$ are defined as follows:

$$\bar{D}^k = \frac{\nabla F\left(\rho^k\right) \rho^k + \rho^k \nabla F\left(\rho^k\right)}{2} - \rho^k \tag{5}$$

$$\tilde{D}^k = \frac{\nabla F\left(\rho^k\right) \rho^k \nabla F\left(\rho^k\right)}{tr\left(\nabla F\left(\rho^k\right) \rho^k \nabla F\left(\rho^k\right)\right)} - \rho^k \tag{6}$$

$$q(t_k) = 1 + 2t_k + t_k^2 tr\left(\nabla F\left(\rho^k\right)\rho^k \nabla F\left(\rho^k\right)\right) \tag{7}$$

where $t_k \in [0, 1]$, $q(t_k) \geq 1$. To show this algorithm's robustness, we randomly initialize the diagonal matrix ρ^0 while it satisfies $\rho^0 > 0$ and $tr\left(\rho^0\right) = 1$.

At the k-th iteration, after generating an ascent direction D^k, ρ is updated as follows:

$$\rho^{k+1} = \rho^k + t_k D^k \tag{8}$$

where t_k is the step length. This process will stop when the change in the objective function $F(\rho)$ is less than a threshold ϵ. We set $\epsilon = 10^{-5}$ empirically in this paper. We observed that the convergence speed is slowing down and the value of objective function is beginning to stabilize when the change in the objective function is less than the threshold. The complete procedure of density matrix estimation is described in Algorithm 1.

Algorithm 1 Algorithm of estimating density matrix for text and image

Require: Each (visual) word vector (s_i) w_i, the initial density matrix ρ^0 and each document d
Ensure: Density matrix of each document ρ

// Constructing the projector
$\mathcal{P}_U \Leftarrow \phi$; // $\mathcal{P}_U$ is the projector sequence
for each $d \in D$ **do**
  for each $w \in d$ **do**
    //w is a single term or a visual word
    for $i = 1; i \leq \#(w, d); i++$ **do**
      $\Pi_i = |w_i\rangle\langle w_i|$;
      $\mathcal{P}_U \Leftarrow \mathcal{P}_U \bigoplus \Pi_i$; //add the projector to the sequence
    end for
  end for
end for
// Train density matices ρ
for each $\mathcal{P}_U$ of d **do**
  Maximize $F(\rho) \equiv \sum_i \log\left(tr\left(\Pi_i \rho\right)\right)$;
  for $k = 1; F\left(\rho^{k+1}\right) - F\left(\rho^k\right) \leq \epsilon = 10^{-5}; k++$ **do**
    $\rho^{k+1} = \rho^k + t_k D^k$;
  end for
end for
return ρ

Finally, the Dirichlet smoothing method is applied to smooth the density matrices. Let ρ_{doc} be a document QMR obtained by MLE, it is then smoothed by interpolation with the collection QMR ρ_{col}:

$$\rho_d = (1 - \gamma)\rho_{doc} + \gamma\rho_{col} \tag{9}$$

where $\gamma \in [0, 1]$ controls the amount of smoothing. $\gamma = \frac{\mu}{\mu + M}$ is a commonly used form of the parameter for Dirichlet smoothing [65]. In our work, μ is a parameter. M is the number of quantum events occuring in the collection.

For a clearer illustration, we give an example to interpret the whole process of calculating the density matrices, as follows.

Example: Consider a textual document: "the dog and the music".

1. We pre-process the text by removing stop words, resulting in the pre-processed text "dog and music".

2. We use the glove tool to find each word's embedding, with the dimensionality 3. For example, we have w(dog) = (-0.15, -0.24, 0.31), w(and) = (0.36, 0.86, -0.61), and similarily, w(music) = (-0.92, 0.59, 0.43).

3. Performing vector normalization: $w(dog) = \frac{w(dog)}{|w(dog)|}$ = (-0.36, -0.57, 0.74), $w(and) = \frac{w(and)}{|w(and)|}$ = (0.32, 0.77, -0.55), $w(music) = \frac{w(music)}{|w(music)|}$ = (-0.78, 0.50, 0.37).

3. After normalization, we can construct each word's projector using Eq.1. Specifically, we have: $\Pi_{dog} = w(dog)^T \cdot w(dog) = \begin{bmatrix} 0.13 & 0.21 & -0.27 \\ 0.21 & 0.32 & -0.42 \\ -0.27 & -0.42 & 0.55 \end{bmatrix}$, $\Pi_{and} = w(and)^T \cdot w(and) = \begin{bmatrix} 0.10 & 0.25 & -0.18 \\ 0.25 & 0.59 & -0.42 \\ -0.18 & -0.42 & 0.30 \end{bmatrix}$, $\Pi_{music} = w(music)^T \cdot w(music) = \begin{bmatrix} 0.61 & -0.39 & -0.29 \\ -0.39 & 0.25 & 0.19 \\ -0.29 & 0.19 & 0.14 \end{bmatrix}$.

4. Then each projector is a 3*3 matrix. The textual document is now be represented as a sequence of projectors. In this example, the document is comprised of three projectors: $\mathcal{P}_U = \left\{\Pi_{dog}, \Pi_{and}, \Pi_{music}\right\}$.

5. Based on quantum probability theory, the probability of each word is $p(word) = tr(|w_i\rangle\langle w_i|\rho)$. We can calculate each word's probability, given a random initial $\rho_0 = \begin{bmatrix} 0.46 & 0 & 0 \\ 0 & 0.49 & 0 \\ 0 & 0 & 0.05 \end{bmatrix}$. Hence, $p(dog)$= 0.24, $p(and)$= 0.35, $p(music)$= 0.41. The probability of the document is the product of probabilities of all words, i.e., $p(doc) = p(dog) \cdot p(and) \cdot p(music)$ = 0.03.

6. We use the Maximum Likelihood Estimation to train the final density matrix. In our work, we employ the globally convergent algorithm to maximize the objective function $F(\rho)$. Giving the ρ_0, we can calculate $F\left(\rho^0\right)$ = -5.47. Then we will update the density matrix using the Eq.8 (i.e., $\rho^1 = \rho^0 + t_0 D^0 = \begin{bmatrix} 0.32 & 0.09 & -0.16 \\ 0.09 & 0.51 & -0.21 \\ -0.16 & -0.21 & 0.17 \end{bmatrix}$), and check if $F\left(\rho^{k+1}\right) - F\left(\rho^k\right) \geq 10^{-5}$ at each iteration. If $F\left(\rho^{k+1}\right) - F\left(\rho^k\right) < 10^{-5}$, we will tune t dynamically to get a new D, and then calculate the objective function again.

7. Finally, we get the 3*3 density matrix representation of this document: $\begin{bmatrix} 0.41 & 0.03 & -0.07 \\ 0.03 & 0.52 & -0.08 \\ -0.07 & -0.08 & 0.07 \end{bmatrix}$.

In this way, we have obtained density matrices that represent text and images respectively. We will identify the overall sentiment of multimodal data using these matrices and a multimodal decision fusion strategy, which will be detailed in the next section.

4. A Quantum Interference-inspired Multimodal Decision Fusion (QIMF) Strategy

In this section, we first introduce the double-slit experiment and the Quantum Interference effect (QI). We then elaborate its analogy to multimodal sentiment analysis. Finally, we propose a Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy.

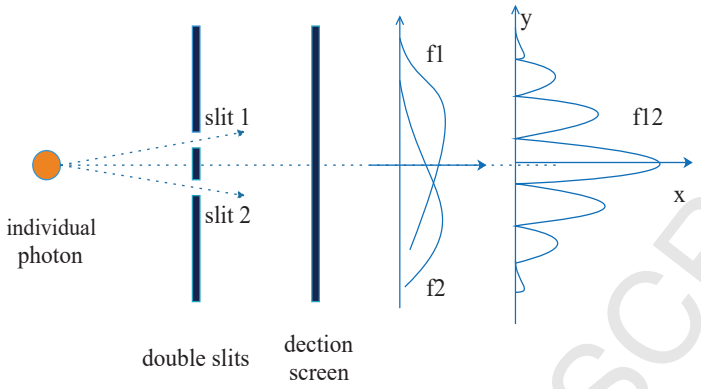


Figure 2: The double-slit experiment. f_1 (*or* f_2) is the curve observed by closing slit 2 (or slit 1). f_{12} is the curve observed by opening both slit 1 and slit 2. Clearly, $f_{12} \neq f_1 + f_2$ because of the interference effect.

4.1. The Double-slit Experiment

The double-slit interference experiment [66], as shown in Figure 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. If the photon passes through just one slit, then it cannot pass through the other slit to create an interference pattern. This strange behavior of microscopic particles demonstrating such a quantum interference effect cannot be explained sufficiently with any classical theory. Hence, it is necessary to introduce QT to interpret the behavior.

In QT, the wave function $\varphi(x)$ [67] is a probability amplitude function of position x, which is a complex number. It is a good description of the quantum state of a particle and can be 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:

$$\varphi_p(x) = \alpha\varphi_1(x) + \beta\varphi_2(x) \tag{10}$$

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

$P(x) = |\varphi(x)|^2$ determines the probability (density) that a particle in the 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. Therefore, the curves f_1 and f_2 are measured as:

$$f_1 = |\alpha|^2 |\varphi_1(x)|^2 \tag{11}$$

$$f_2 = |\beta|^2 |\varphi_2(x)|^2 \tag{12}$$

We have that the probability distribution f_{12}:

$$\begin{aligned} f_{12}(x) &= \left|\varphi_p(x)\right|^2 = |\alpha\varphi_1(x) + \beta\varphi_2(x)|^2 \\ &= (\alpha\varphi_1(x) + \beta\varphi_2(x)) \cdot (\alpha\varphi_1(x) + \beta\varphi_2(x))^\dagger \\ &= \alpha\varphi_1(x) \cdot (\alpha\varphi_1(x))^\dagger + \beta\varphi_2(x) \cdot (\beta\varphi_2(x))^\dagger \\ &+ \alpha\varphi_1(x) \cdot (\beta\varphi_2(x))^\dagger + \beta\varphi_2(x) \cdot (\alpha\varphi_1(x))^\dagger \\ &= \alpha\varphi_1(x) \cdot (\alpha\varphi_1(x))^\dagger + \beta\varphi_2(x) \cdot (\beta\varphi_2(x))^\dagger \\ &+ \alpha\varphi_1(x) \cdot (\beta\varphi_2(x))^\dagger + \left(\alpha\varphi_1(x) \cdot (\beta\varphi_2(x))^\dagger\right)^\dagger \\ &= |\alpha\varphi_1(x)|^2 + |\beta\varphi_2(x)|^2 + 2Re\left(\alpha\varphi_1(x) \cdot (\beta\varphi_2(x))^\dagger\right) \\ &= |\alpha\varphi_1(x)|^2 + |\beta\varphi_2(x)|^2 + 2\,|\alpha\varphi_1(x)\beta\varphi_2(x)|\,cos\theta \\ &= f_1 + f_2 + 2\sqrt{f_1 f_2}cos\theta \end{aligned} \tag{13}$$

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

4.2. An Analogy in Multimodal Sentiment Analysis

We draw an analogy to the double-slit experiment in multimodal sentiment analysis. The sentiment label of multimodal document is uncertain, which can be analogized as the photon. The sentiment of 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 Figure 3. In our analogy, the sentiment information of each modality will influence the final decision simultaneously. If the sentiment of the text and the image both are +1 (or -1), then the final sentiment score most certainly is +2 (very positive) (or -2, very negative). This phenomenon can be viewed as the constructive interference. Note that we elaborate this analogy for developing a new multimodal fusion strategy, instead of modeling the psychological process. In this paper, we believe that the mathematical equations used to describe quantum interference also serve as handy information fusion rules.

We use the wave function $\varphi(x)$ to formalize our analogy. The sentiment polarity of multimodal documents can be analogized as a combination of the sentiment of the text and the image, as shown below:

$$\varphi_u(x) = \alpha\varphi_t(x) + \beta\varphi_i(x) \tag{14}$$

where $\varphi_t(x)$ is the wave function of the sentiment of the text, $\varphi_i(x)$ is the wave function of the sentiment of the image. Therefore, the probability distribution of the sentiment polarity of the text or the image can be respectively formulated as:

$$f_t = |\alpha|^2\,|\varphi_t(x)|^2 \tag{15}$$

$$f_i = |\beta|^2\,|\varphi_i(x)|^2 \tag{16}$$

The probability distribution of the final sentiment score can be measured as:

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

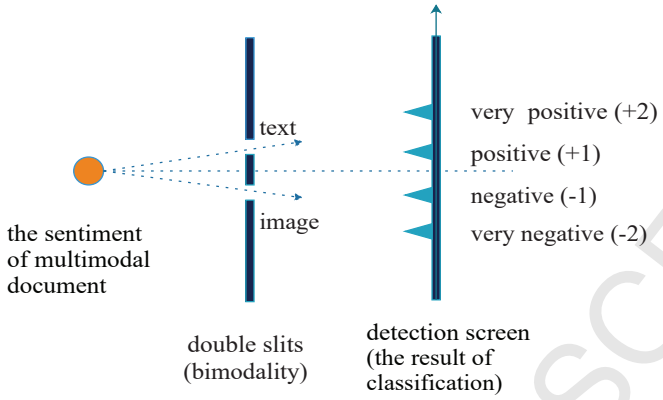


Figure 3: Our analogy with multimodal sentiment analysis and the double-slit experiment.

Note that in this paper we do not explicitly use complex numbers, but the framework is general. Indeed, the complex number is real if its imaginary part is zero, and thus the set of real numbers is a proper subset of the set of complex numbers. Busemeyer and Bruza [23] have mentioned that the beauty of complex numbers is that the formulation retains their simplicity and elegance no matter whether they are based on real or complex numbers. Therefore, quantum interference theory can be used to help develop a novel multimodal sentiment analysis model.

4.3. Developing the QIMF Strategy

Based on the above analogy, we can explore the correlation among multi-modalities at the decision level, and propose a Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy. Compared with the previous decision fusion strategies, QIMF adds an interference term.

$P(x) = |\varphi(x)|^2$ describes the probability of position x. At the decision level, we can refer to $P(x)$ as the probability of the sentiment score x (x=+1,+2,-1,-2). Similarly, we interpret $P_t(x) = |\varphi_t(x)|^2$ as the probability that the sentiment score of the text is x, denoted as P_t. We interpret $P_i(x) = |\varphi_i(x)|^2$ as the probability that the sentiment score of the image is x, denoted as P_i. The final decision P_u can be written as:

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

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

5. Our Quantum-inspired Multimodal Sentiment Analysis (QMSA) Framework

We have described the Quantum-inspired Multimodal Representation (QMR) model (in Section 3.2) and the Quantum Interference inspired Multimodal decision Fusion (QIMF) strategy (in Section 4.3), respectively. Now we introduce our Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework.

Algorithm 2 Framework for Quantum-inspired Multimodal Sentiment Analysis (QMSA)

Input: Multimodal documents (the text d_{text} and the image d_{image})
Output: Positive or negative sentiment labels (+1, -1)

1: Map each term of the text d_{text} to vector w_i using word embedding; Map each visual word of the image d_{image} to vector s_i using bag of features model.
2: Input w_i and s_i, and utilize the QMR model to train density matrices ρ_{text} and ρ_{image}. Use these density matrices ρ_{text} and ρ_{image} to represent the text and the image, refer to Algorithm 1.
3: Employ density matrices ρ_{text} and ρ_{image} as input data, train their own classifiers C_{text} and C_{image}, and get the decision results P_{text} and P_{image}, respectively.
4: Perform the QIMF strategy to fuse the decisions P_{text}, P_{image}, tune the weights α, β and $\cos\theta$. We define that the weights of $\cos\theta$ can range from -1 to 0 to +1, with a default interval of 0.1. The weights of α, β must satisfy $|\alpha|^2+|\beta|^2 = 1$. Then get the final decision of multimodal sentiment analysis P_u.

The QMSA framework adopts the QMR model to represent the text and the image, which are encapsulated in density matrices ρ_{text} and ρ_{image}. Using density matrices as input data and choosing appropriate classifiers (denoted C_{text}, C_{image}) would lead to their own decisions, denoted P_{text}, P_{image}. The QMSA framework then applies the QIMF strategy (Equation 20) to fuse the decisions. Finally, the QMSA framework produces a final decision of multimodal sentiment analysis P_u through tuning the weights α, β and $\cos\theta$. This framework is described in Algorithm 2.

6. Experiments

In this section, we conduct extensive experiments to evaluate the performance of our QMSA framework, including the performance of the QMR model and QIMF strategy individually and as a whole.

6.1. Experimental Settings

As a matter of fact, currently there is a lack of large scale, open-access and well labeled datasets for multimodal sentiment analysis. We create two large datasets to support our experiment. As what other researchers did in [50, 56, 8, 68], first, we set a list of keywords with strongly positive and negative sentiment using SentiWordNet (which is a well-known dictionary)[34]. Then, we query Flickr and Getty Images with these words, and use the labels of these words to label the retrieved images of the first ten pages.

As a result, we have gathered 99,351 multimodal documents from Flickr and 171,793 from Getty Images respectively, using 127 keywords. We have made our datasets freely downloadable [1]. Table 1 shows some statistics of our collected multimodal datasets. In our work, we mainly use two software: Matlab 2014 a/b and Python 2.7, which are installed on Windows 8.1 and Windows server 2012, to support our implementation.

The multimodal data are pre-processed as follows. The overly large images (i.e., size exceeding 1000 pixel*1000 pixel) are re-sized. For text, we remove the stop words and punctuations

[1]Both datasets can be accessed on the web page: http://www.tjucs.win/faculty/dsong/yazhouzhang.html, https://pan.baidu.com/s/1bqoscfP

Table 1: Our dataset from GI and Flickr

Sentiment	Num of keywords	Num of GI	Num of Flickr
Positive	62	91,419	49,728
Negative	65	80,374	49,623
Sum	127	171,793	99,351

using a standard stopword list. We employ the 10-fold cross-validation method to evaluate all models in this paper.

In this paper, we use the Glove tool to produce word embeddings. It is worth mentioning that we also train our embeddings using the Gensim API [69], and we find that it makes no difference from the Glove. The dimensionality is set to 100 instead of 300, considering the computation cost for classification. We believe that a 100-dimensional vector has embodied sufficiently rich semantic information. Similarly, we set the dimensionality of visual words to 128, which is the default setting of the SIFT algorithm. Under these settings, the QMR model is implemented.

In order to demonstrate the robustness of our proposed QMSA framework and show the impact of different classification algorithms, we choose two representative classifiers, Random Forest (RF) and Support Vector Machines (SVM) [70], which have been considered as the state of the art when dealing with the sentiment analysis problems [46][71]. For RF, we set the number of trees in the forest to 500 and other weights of RF as the default values, e.g., "bootstrap" as "True", "criterion" as "gini", etc. For SVM, we set the kernel function to "linear" because of the large scale data and features, and "probability" to "True". Other weights in SVM are set as the default values, e.g., "coef0 " as zero, "gamma" as "auto", etc. Therefore, the experimental results can be easily replicated.

Then, we perform our QIMF strategy by using the prediction probability of the sentiment scores (+1, -1) for each sample data. For example, assume that the prediction probability of +1 for a text document is 0.6, i.e., $p_t\,(x = +1) = 0.6$, and that the prediction probability of +1 for an image document is 0.4, i.e., $p_i\,(x = +1) = 0.4$. We can get the multimodal fusion decision result $p_u\,(x = \pm 1)$ by tuning the parameters α, β and $\cos\theta$ (Equation 20).

We compare the performance of our proposed models with a wide range of baseline algorithms as follows. We adopt **Precision**, **Recall**, **F1 score**, **Accuracy** and **ROC curve** as evaluation metrics to measure the classification performance of each method with two classes, positive and negative. We employ t-test to perform the significance test in this experiment.

Single visual model: we use bag of visual words method (bovw) [72] to generate histograms of visual word occurrences that represent images, and train a Random Forest (RF) classifier or an SVM classifier (whose parameters use the same settings as above) to analyze the polarity of the images in the testing set.

Single textual model: we use word embeddings, for which the dimensionality is set to 100 [73], to represent all textual documents, and train a RF classifier or an SVM classifier (whose parameters use the same settings as above) to analyze the polarity of the text.

Bag of words model: we use the classical bag of words method (bow) to generate histograms of word frequencies that represent texts, and train a RF classifier or an SVM classifier (whose parameters use the same settings as above) to analyze the polarity of the texts in the testing set. To limit the size of the feature vectors, we use the 3000 most frequent words. We filter out the stop words using a standard english stop word list, which is encapsulated in the NLTK tool [74].

Table 2: One example of the mass function and the joint mass function on GI dataset

Hypothesis	$Mass_{text}$	$Mass_{image}$	K	$Mass_{multimodal}$
+1	0.5251	0.3598	0.4930	0.3832
-1	0.4749	0.6402	0.4930	0.6167

Feature-level Multimodal Fusion model (FMF): we concatenate 128-dimensional visual vector and 100-dimensional textual vector at the feature level, and then train a RF classifier or an SVM classifier (whose parameters use the same settings as above) to identify the overall sentimental polarity of multimodal documents in the testing set.

Majority Voting Fusion model (MVF): based on visual vector (which is extracted by bag of visual word method) and textual vector (which is extracted by the Gensim API), we train two RF models or two SVM models to get the local decisions, respectively. We combine the local decisions using the popular rule-based decision fusion strategy: majority voting [75].

Linear Weighted Fusion model (LWF): based on visual vector (which is extracted by bag of visual words method) and textual vector (which is extracted by the Gensim API), we train two RF models or two SVM models to get the local decisions, respectively. We combine the local decisions using a linear weighted fusion strategy. We assign different weights to different modalities, which refers to $P_u = \omega_1 P_{text} + \omega_2 P_{image} + \omega_3 P_{text} P_{image}$, where $\omega_1 \in [0, 1]$, $\omega_2 \in [0, 1]$, $\omega_3 \in [-1, 1]$. Note that we make a relaxation that ω_1 plus ω_2 does not necessarily equal to one.

Dempster-Shafer Evidence Fusion model (DSEF): 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 [76]. In this paper, each visual vector (which is extracted by bag of visual words method) and textual vector (which is extracted by the Gensim API) are issued to the classifiers, returning two result lists with different probability scores. Hence, two sentiment scores (which are +1,-1) construct the power set. We use the probability scores, which are offered by the classifiers, to specify the mass function. According to the D-S theory, the combination (called the joint mass) is calculated from the two sets of masses m_{text} and m_{image} in the following manner: $m_{multimodal}(A) = \left(m_{text} \oplus m_{image}\right)(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_{text}(B) m_{image}(C)$, where $K = \sum_{B \cap C = \varnothing} m_{text}(B)\, m_{image}(C)$. Table 2 shows an example of the mass function and the joint mass function.

Multimodal Deep Learning model (MDL): considering the popularity of deep learning, we can learn a joint representation for various features extracted in different modalities, which is similar to [77]. In [77], the authors used Restricted Boltzmann Machine (RBM) to learn the joint distribution over image and text inputs. We choose to replace RBM with Convolutional Neural Networks (CNN) to learn the joint distribution over image and text inputs through constructing a shared hidden layer based on the similar framework. The MDL model uses a feature-level fusion strategy.

Deep Convolutional Neural Networks (DCNN): we also compare our framework with visual and textual sentiment analysis model using Deep Convolutional Neural Networks [57]. We first train a CNN on top of word vectors for textual sentiment analysis and employe a CNN for visual sentiment analysis, then use Logistic Regression to perform sentiment prediction of the text and the image individually. Finally, we fuse the probabilistic results using the average strategy. We set the dimensionality of word embeddings to 100 and resize the images to 64×64, for consistency with the work reported in the original paper [57]. The DCNN model uses a decision-level

fusion strategy.

QLM: in order to validate the effectiveness of the globally convergent algorithm, we compare our framework with the QLM, which uses the original $R\rho R$ algorithm to estimate the maximum likelihood value.

Our proposed models are listed below:

QIMF model: since we get the local decisions from **Single visual model** and **Single text model**, respectively, we can use the QIMF strategy (in Section 4.3) to make the final decision through tuning different α, β parameters.

QMR model: we use the QMR model (in Section 3.2) to represent the image and the text separately, and concatenate both visual and textual features as the multimodal feature. Then, we perform the sentiment recognition using a RF or an SVM classifier, whose parameters use the same settings as above. This model aims at validating whether the QMR model could fill the "semantic gap" between low-level features and high-level semantic labels.

Q-LWF framework (QMR+LWF): we first adopt the QMR mdoel to represent the image and the text separately, and get their own local decisions using a RF or an SVM classifier. Second, we perform a linear weighted fusion strategy to obtain the final results through tuning $\omega_1, \omega_2, \omega_3$ parameters. We construct this framework to compare with the QMSA framework, aiming to demonstrate the effectiveness of our quantum-interference inspired decision fusion strategy.

Q-DSEF framework (QMR+DSEF): we first adopt the QMR mdoel to represent the image and the text separately, and get their own local decisions, i.e., the probability scores from a RF or an SVM classifier. Second, we perform the Dempster-Shafer evidence theory to fuse the final results. We define the mass function similar to the work in [78]: $m(\{md_i\}) = m_{text}(\{md_i\}) \times m_{image}(\{md_i\}) + m_{text}(\{\Theta\}) \times m_{image}(\{md_i\}) + m_{text}(\{md_i\}) \times m_{image}(\{\Theta\})$, where $m_k(\{md_i\})$ (which k = text, image) can be considered as the probability that the sentiment of the multimodal document md_i is +1 or -1. Θ denotes the whole dataset, and $m(\Theta)$ represents the uncertainty in those sources of evidence. In this paper, $m(\Theta)$ is defined as: $m_{k(\Theta)} = 1 - \frac{\sum_{i=1}^{N} m_k(\{md_i\})}{\sum_{i=1}^{N} m_{text}(\{md_i\}) + \sum_{i=1}^{N} m_{image}(\{md_i\})}$. N is the number of the multimodal documents. We construct this framework to compare with the QMSA framework, aiming to demonstrate the effectiveness of our quantum-interference inspired decision fusion strategy.

QMSA framework (QMR+QIMF): we first adopt the QMR mdoel to represent the image and the text separately, and get their own local decisions using a RF or an SVM classifier. Secondly, we perform the QIMF strategy to obtain the final result through tuning different α, β parameters. This framework validate whether our framework could deal with the multimodal sentiment analysis task. We have made our codes of QMSA framework open-source for free download[2].

6.2. Results on Getty Images (GI) Dataset

The first set of experiments are conducted on the Getty Images dataset, where the text description of multimodal documents is generally more formal than that in other social applications. Table 4 shows the performance of different approaches using two classifiers on Getty Images datasets.

First, we analyze the experimental results of using RF classifier. From Table 4, it is observed that single visual model performs poorly. This result indicates that it is insufficient

[2]The source codes of our QMSA framework including text and image sentiment analysis can be accessed on the web page: http://www.tjucs.win/faculty/dsong/yazhouzhang.html.

to only utilize low-level visual features to analyze the sentiment polarity of images. Compared with single visual model, single textual model improves the performance as we expected. As a widely used model, the bag of words model gets the highest precision and accuracy results among all baselines. This implies that the bag of words model can extract textual feature better than single textual model. However, we find that the bag of words model relies on the classifiers through comparing the results of using RF and SVM. Through concatenating textual features and visual features, the FMF model produces lower performances than single visual model and single textual model. This shows that simple concatenation strategy is not able to capture the correlation between multi-modalities, and may also bring noise in feature representation. One should be careful to adopt this strategy for multimodal sentiment analysis.

A comparison with three baseline decision fusion strategies: the MVF model uses the majority voting strategy to fuse the local decisions from Single visual model and Single textual model, and gets good results. As one of the most popular decision-level fusion algorithms, the majority voting strategy chooses to trust in the highest decision score. The LWF model uses a weighted fusion strategy, which relaxes the constraint on the coefficients so that ω_1 plus ω_2 does not necessarily equal to one. When $\omega_1 = 0.5$, $\omega_2 = 0.1$ and $\omega_3 = 0.3$, the LWF model gets its highest classification scores. We can observe that the LWF model outperforms other baselines, which means a relaxed linear weighted fusion strategy can more effectively incorporate some complementary decision information offered by different modalities. However, this fusion strategy is different from the linear combination strategy, because it optionally relaxes some constraint conditions. As a general framework for reasoning with uncertainty, the Dempster-Shafer (D-S) evidence theory is also taken as a baseline. It gets the lowest classification results among these three strategies. We think that this baseline largely relies on how to define the mass function and the judgement rule. We use an elaborated method to define the mass function in the Q-DSEF framework, which will be discussed later this this section.

The DCNN model outperforms the MDL model, which indicates that training two models separately is better than training a joint representation model when using deep learning techniques. The QLM gets good results, which demonstrates that applying quantum probability theory is flexible for developing novel sentiment analysis models.

Compared with the above baselines, the QIMF model shows a better performance over the FMF, MVF, DSEF and single modality models. Unsurprisingly, the QIMF model achieves the same accuracy result, in comparison with the LWF model. Because that the LWF model relaxes a few constraint conditions. It indicates that our proposed decision fusion strategy is an effective decision fusion strategy, which has its mathematical principle. Compared with the MVF model, the accuracy result has increased by about 2%. The performances obtained by our QMR model illustrate the benefits of using density matrices, which are probability distributions of events (words). Compared with the QLM, the accuracy result has increased by about 5%, which demonstrates the effectiveness of the globally convergent algorithm. Compared with all baselines, we believe that density matrix can carry more semantic information than vector-based representation models. This result demonstrates that our proposed QMR model is an effective representation learning model.

A comparison of three multimodal sentiment analysis frameworks: We have combined the QMR model with the LWF fusion strategy, and tune free parameters $\omega_1, \omega_2, \omega_3$. When $\omega_1 = 0.9$, $\omega_2 = 0.5$ and $\omega_3 = -0.1$, the Q-LWF framework gets its highest classification results. Meanwhile, it also achieves the best performance on precision and accuracy metrics. Compared with the LWF model, the accuracy result has increased by about 4%, which demonstrates the effectiveness of our quantum-inspired representation model. Moreover, we have combined the

QMR model with the D-S evidence theory, and proposed a complex mass function. This mass function considers the uncertainty in multimodal sentiment analysis. We can observe that the Q-DSEF framework obtains better performance than the DSEF framework. Finally, we have also combined the QMR model with the QIMF classification strategy, and test its performance on the dataset. We tune free parameters α, β to make $\alpha^2 = 0.7, \beta^2 = 0.3$, which means we would pay more attention on the text. When $\cos\theta = 0.3$, our QMSA framework achieves the best performance on recall, f-score and accuracy metrics. The accuracy result of our framework is over 88%, with an improvement of about 4% over the LWF model (which gets the highest results among all baselines). Compared with the QIMF model and the QMR model, the accuracy of the QMSA framework increases by about 4% and 3%, respectively. Compared with the Q-LWF framework, the QMSA framework gets higher recall and f1 classification results, and get the same accuracy result. Compared with the Q-DSEF framework, the QMSA framework achieves better performance on all metrics. This implies that the quantum interference inspired decision fusion strategy is an effective fusion strategy, which is also rooted on a well-founded mathematical derivation. Overall, we attribute the improvements to both QMR model and QIMF strategy. It suggests that: a) an effective semantic learning model could help the machine to better "understand" multimodal documents; b) the QIMF strategy indeed incorporate some complementary decision information.

Now, we analyze the results of using SVM classifier. Overall, we can observe that the performance of all models using SVM is not as good as that using RF. From the perspective of classifier, RF is often claimed to be better at dealing with super large scale of training samples. From the perspective of dataset, since the Getty Images dataset is crawled from Getty Images some images contain digital watermark. However, the watermark is tolerable due to the relatively formal and clean descriptions of multimodal documents as argued in [68]. Moreover, since we run all models on the same dataset, the impact of watermark applies to all models. Because of these reasons, RF and SVM give different classification results.

Nevertheless, from the SVM results, we can still get the similar observations as with the RF results. The QIMF model outperforms single visual model, single textual model, the MVF model and the DSEF model. Compared with the LWF model, the QIMF model achieves the same accuracy result but higher recall and f1 results. Our QIMF model could access different information from textual and visual models via different weights. This shows that introducing an effective and principled decision-level fusion strategy is better than unimodal sentiment analysis model and a simple fusion strategy. It is worth noting that the DSEF model and the Q-DSEF framework perform poorly. we analyze our data and probability scores, and think that the D-S evidence theory relies on the classifiers. When different classifiers give the opposite classification results, the D-S evidence theory may exacerbate uncertainty about the decision. The QMR model produces a very large improvement over the MVF model, about 24%. In addition, we tune free parameters α, β to obtain $\alpha^2 = 0.8, \beta^2 = 0.2$. When $\cos\theta = -0.6$, our QMSA framework gets a 73.76% accuracy, which outperforms the Single visual, Single textual, bag of words, FMF, MVF, DSEF, MDL and QLM models. It indicates that extracting more features at feature level plus accessing more information at decision level lead to a better performance. In order to provide a clearer empirical sense of our model's performance, a complete ROC curve is shown in Figure 4. Moreover, a series of positive and negative examples (for which our framework makes sound judgment while the MVF and FMF models classify them wrongly) from the Getty Images dataset are shown in Figure 8 for illustration.

The computational time: Because both datasets are very large and the training and testing samples for the quantum-inspired models are all matrices, the computation time used for training

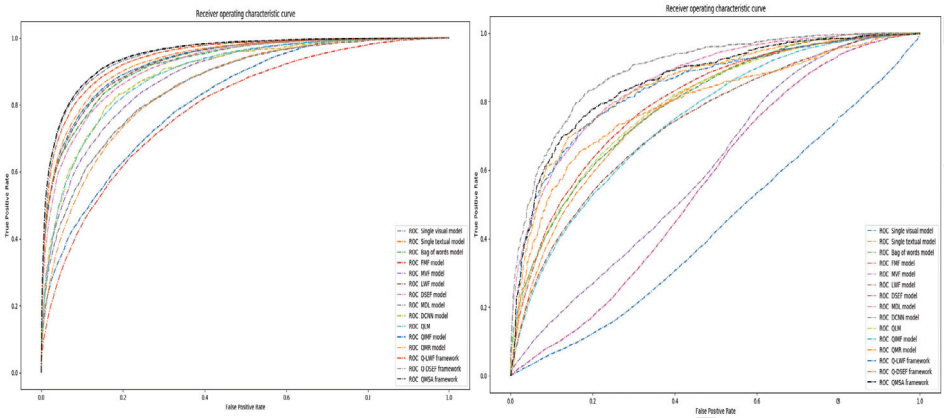


Figure 4: the complete ROC curves on Getty Images dataset. (a) ROC curves of RF classifier; (b) ROC curves of SVM classifier.

Table 3: The computational time of most models on GI dataset.

Classifier	Model	Computational time (h)	Model	Computational time (h)
RF	QMSA	144.5	Q-DSEF	145.0
	Q-LWF	144.5	QMR	256.7
	FMF	103.8	DSEF	42.2
	MVF	46.6	LWF	46.6
	Single visual model	39.7	Single textual model	12.4
SVM	QMSA	696.0	Q-DSEF	696.4
	Q-LWF	696.0	QMR	893.3
	FMF	233.0	DSEF	154.4
	MVF	194.2	LWF	194.2
	Single visual model	143.1	Single textual model	38.5

and classification is longer than the use of other baselines. Random Forest has a better ability to deal with large scale of samples. The total computation time of QMSA framework using RF classifier to complete all 10-fold validation experiments on the Getty Images dataset is almost 144.5 hours, longer than that of the FMF model (103.8 hours), DSEF model (42.2 hours) and MVF model (46.6 hours). The computational time of QMSA framework using SVM classifier is almost 696 hours, also longer than that of the FMF model (233 hours), DSEF model (154.4 hours) and MVF model (194.2 hours). Table 3 summarizes the computational time of these models.

Table 4: The Performance on Getty Images. Best results are highlighted in boldface. Numbers in parentheses indicate relative improvement over the MVF model. The symbol † means statistical improvement over all baselines.

Classifier	Algorithm	Precision	Recall	F1	Accuracy
RF	Single visual model	0.7247	0.8073	0.7638	0.7341
	Single textual model	0.7729	0.8125	0.7919	0.7739
	Bag of words model	0.8622	0.8163	0.8400	0.8304
	FMF model	0.7061	0.8079	0.7536	0.7180
	MVF model	0.8105	0.8355	0.8226	0.8278
	LWF model	0.8478	0.8336	0.8360	0.8459
	DSEF model	0.7928	0.9104	0.8475	0.8252
	MDL model	0.7844	0.8137	0.7981	0.7912
	DCNN model	0.8457	0.7829	0.8132	0.8111
	QLM	0.8289	0.8519	0.8393	0.8210
	QIMF model	0.8165 (+0.74%)	0.9072 (+8.58%)	0.8604 (+4.59%)	0.8459 (+2.18%)
	QMR model	0.8652† (+6.75%)	0.8779† (+5.07%)	0.8715† (+5.94%)	0.8615† (+4.07%)
	Q-LWF framework	**0.8828**† (+8.24%)	0.8708† (+4.46%)	0.8745† (+6.30%)	**0.8824**† (+6.60%)
	Q-DSEF framework	0.8637† (+6.56%)	0.9116† (+11.29%)	0.8870† (+9.11%)	0.8758† (+5.79%)
	QMSA framework	0.8794† (+8.50%)	**0.9152**† (+9.54%)	**0.8969**† (+9.03%)	**0.8824**† (+6.60%)
SVM	Single visual model	0.6002	0.8633	0.7123	0.6076
	Single textual model	0.6830	0.8267	0.7480	0.7035
	Bag of words model	0.7327	0.7554	0.7435	0.7212
	FMF model	0.7298	0.8143	0.7697	0.7400
	MVF model	0.6261	0.8111	0.7083	0.6307
	LWF model	0.7173	0.7156	0.7168	0.7142
	DSEF model	0.5865	0.7858	0.6717	0.5915
	MDL model	0.7844	0.8137	0.7981	0.7912
	DCNN model	**0.8457**	0.7829	**0.8132**	**0.8111**
	QLM	0.7329	0.7161	0.7244	0.7296
	QIMF model	0.6726 (+7.42%)	**0.8739** (+7.74%)	0.7662 (+6.77%)	0.7142 (+13.24%)
	QMR model	0.7561† (+20.76%)	0.8413† (+3.72%)	0.7951† (+12.25%)	0.7808† (+23.79%)
	Q-LWF framework	0.7884† (+25.92%)	0.7940† (-2.10%)	0.7900† (+11.53%)	0.7976† (+26.46%)
	Q-DSEF framework	0.6564† (+4.84%)	0.8270† (+1.96%)	0.7318† (+3.17%)	0.6961† (+10.36%)
	QMSA framework	0.8034† (+28.31%)	0.7785† (-4.02%)	0.7912† (+11.70%)	0.7976† (+26.46%)

6.3. Results on Flickr Dataset

We conduct the second set of experiments on the Flickr dataset. Since multimodal documents from Flickr are more diverse and informal, analyzing the sentiment of Flickr data is considered more more challenging. Table 5 shows the results.

First, we analyze the experimental results of using RF classifier. We can see that single visual model has the worst performance, and single textual model can achieve a higher accuracy but a lower precision. These results indicate again that sentiment recognition from images is not as effective as that from text. The bag of words model gets the second highest accuracy result among baselines. However, it relies on the classifiers. We observe that the performance

declines sharply when using SVM. The FMF model produces a modest improvement over mono-modality models (single textual model and single visual model). This implies that multimodal content do express more accurate sentiment than mono-modal content. It also suggests that simple concatenation strategy is not enough to deal with multimodal tasks, and it is necessary to explore effective feature representation methods.

A comparison of three decision fusion strategies: The MVF model adopts a majority voting strategy to fuse the local decisions, and also produces modest improvement. The LWF model and the DSEF model outperform the MVF model, which implies that it is helpful to develop more refined decision fusion strategy. When $\omega_1 = 0.8$, $\omega_2 = 0.3$ and $\omega_3 = 0.3$, the LWF model gets its highest classification scores. As a general framework for reasoning with uncertainty, the DSEF model achieves nearly the same results as the LWF model. It shows that the performance of the D-S evidence theory may depend on the dataset.

Compared with the FMF and MVF models, both deep learning based models (which are the MDL model and the DCNN model) do not achieve very good results. This may be because we do not make a lot of effort in tuning parameters. The QIMF model, which integrates the local decisions from single visual model and single textual model, tends to outperform better than the FMF, MVF, DSEF and single modality models. From the dataset perspective, the text in the Flickr dataset is relatively short and concise, so that text sentiment analysis is in general an easier task than image sentiment analysis. The QIMF model is able to pay more attention on the local decision from textual model through tuning free parameters α, β. The QIMF model gets almost the same classification results as the LWF model. Because the LWF model is a generalization of the QIMF model, through relaxing the original mathematical constraints. However, we aim to propose a novel decision fusion strategy, which is also theoretically more principled.

Our proposed QMR model achieves a noticeable improvement over the above models. Compared with the FMF model, the performance of QMR increases by about 9%. This may be because quantum projectors help to model mid-level term features, and density matrix contain more semantic information. Compared with the QLM, the accuracy result has increased by about 2%, which demonstrates the effectiveness of the globally convergent algorithm. We also tune free parameters α, β to set $\alpha^2 = 0.7, \beta^2 = 0.3$ as in the experiments with Getty images. When $\cos\theta = 0.15$, our QMSA framework significantly outperforms a number of baselines. Through observing the Q-LWF framework, the Q-DSEF framework and the QMSA framework, we believe that extracting mid-level features at feature level plus considering more decision information at decision level lead to the better performance.

Furthermore, we analyze the results of using SVM classifier. We still notice that the performance of SVM is not as good as RF, but get similar observations as in RF. The QIMF model outperforms the single textual, single visual, DSEF model and MVF models. This is because adding an interference term can incorporate some complementary decision information. The QMR model produces a large improvement over the FMF model, about 32%, whereas the perfomance of SVM is very close to the perfomance of RF. This shows that our QMR model does not rely much on classifiers when experimenting on the Flickr dataset. Moreover, we tune the free parameters α, β to obtain $\alpha^2 = 0.8, \beta^2 = 0.2$. When $\cos\theta = -0.9$, our QMSA framework significantly outperforms all baselines. For illustration, a complete ROC curve is shown in Figure 5. Additionally, various positive and negative examples (for which our framework makes sound judgment while the MVF and FMF models mis-classify them) from the Flickr dataset is shown in Figure 9.

The computation time: the computational time of QMSA framework using RF classifier on the Flickr dataset is almost 96 hours, in comparison with that of the FMF model (77.1 hours),

DSEF model (46 hours) and MVF model (46 hours). The computational time of QMSA framework using SVM classifier is almost 504.4 hours, compared with that of the FMF model (119.3 hours), DSEF model (100 hours) and MVF model (which is 100.3 hours). Table 6 summarizes the computational time of these models.

Table 5: The Performance on Flickr. Best results are highlighted in boldface. Numbers in parentheses indicate relative improvement over the MVF model. The symbol † means statistical improvement over all baselines.

Classifier	Algorithm	Precision	Recall	F1	Accuracy
RF	Single visual model	0.6256	0.6450	0.6351	0.6291
	Single textual model	0.7969	0.7989	0.7946	0.7937
	Bag of words model	0.8688	0.9368	0.9020	0.8947
	FMF model	0.8424	0.8467	0.8445	0.8439
	MVF model	0.8267	0.8514	0.8393	0.8358
	LWF model	0.8375	0.8611	0.8491	0.8459
	DSEF model	0.8304	0.8608	0.8453	0.8426
	MDL model	0.8213	0.7999	0.8111	0.8119
	DCNN model	0.8973	0.7167	0.7993	0.8184
	QLM	0.9155	0.9080	0.9169	0.9106
	QIMF model	0.8296 (+0.35%)	0.8533 (+0.22%)	0.8468 (+0.89%)	0.8459 (+1.21%)
	QMR model	0.9278† (+12.23%)	**0.9368**† (+10.03%)	0.9323† (+11.62%)	0.9221† (+10.32%)
	Q-LWF framework	0.9206† (+11.35%)	0.9349† (+9.81%)	0.9275† (+10.58%)	0.9314† (+11.44%)
	Q-DSF framework	0.9170† (+10.92%)	0.9252† (+8.66%)	0.9220† (+9.85%)	0.9251† (+10.68%)
	QMSA framework	**0.9337**† (+12.94%)	0.9288† (+9.09%)	**0.9301**† (+10.82%)	**0.9314**† (+11.44%)
SVM	Single visual model	0.4758	0.7393	0.5770	0.5791
	Single textual model	0.6482	0.6619	0.6592	0.6598
	Bag of words model	0.6759	0.8841	0.7661	0.7367
	FMF model	0.6931	0.7126	0.7027	0.6982
	MVF model	0.6607	0.6933	0.6815	0.6641
	LWF model	0.6864	0.7005	0.6934	0.6917
	DSEF model	0.6343	0.7144	0.6720	0.6516
	MDL model	0.8213	0.7999	0.8111	0.8119
	DCNN model	0.8973	0.7167	0.7993	0.8184
	QLM	0.9009	0.8939	0.8973	0.8982
	QIMF model	0.6884 (+3.25%)	0.7030 (+1.40%)	0.6929 (+1.66%)	0.6917 (+4.56%)
	QMR model	0.9185† (+39.02%)	0.9226† (+33.07%)	0.9215† (+36.68%)	0.9179† (+37.92%)
	Q-LWF framework	0.9174† (+38.85%)	0.9259† (+33.55%)	0.9216† (+35.23%)	**0.9243**† (+39.17%)
	Q-DSF framework	0.9186† (+39.03%)	0.8910† (+28.52%)	0.9156† (+34.35%)	0.9161† (+37.94%)
	QMSA framework	**0.9310**† (+40.91%)	**0.9245**† (+33.35%)	**0.9269**† (+36.00%)	**0.9243**† (+39.17%)

6.4. Remarks on $\cos\theta$

The $\cos\theta$ of the interference term comes from the phase of the product $\alpha\varphi_1(x) \cdot \beta\varphi_2(x)$, which can range from -1 to +1. In this section, we tune $\cos\theta$ with different settings, for a in-

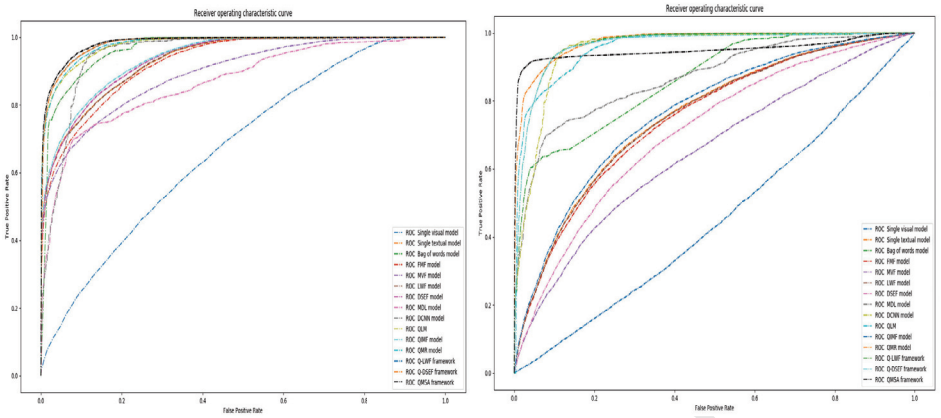


Figure 5: the complete ROC curves on Flickr dataset. (a) ROC curves of RF classifier; (b) ROC curves of SVM classifier.

Table 6: The computational time of most models on Flickr dataset.

Classifier	Model	Computational time (h)	Model	Computational time (h)
RF	QMSA	96.0	Q-DSEF	96.5
	Q-LWF	96.0	QMR	169.7
	FMF	77.1	DSEF	46.0
	MVF	46.0	LWF	46.0
	Single visual model	27.9	Single textual model	10.2
SVM	QMSA	504.4	Q-DSEF	504.9
	Q-LWF	504.4	QMR	700.6
	FMF	119.3	DSEF	100.2
	MVF	100.3	LWF	100.3
	Single visual model	85.5	Single textual model	22.1

depth understanding of the impact of $\cos\theta$. Figure 6 and Figure 7 show the impact of $\cos\theta$ using RF and SVM classifiers respectively.

In Figure 6, we analyze how our QMSA framework behaves on GI and Flickr with respect to the parameter $\cos\theta$ in light of different values of α, β. It is clear that the result increases along with the increase of $\cos\theta$. Specifically, we can observe that the accuracy is highest when $\alpha^2 = 0.7\, and\, \beta^2 = 0.3$ on both datasets. When $\alpha^2 = 0.3\, and\, \beta^2 = 0.7$, the accuracy is the lowest. These two results indicate that analyzing the sentiment of text is more important in multimodal sentiment analysis. When $\alpha^2 = 0.5\, and\, \beta^2 = 0.5$, the accuracy increases until $\cos\theta = -0.4$, and then keeps unchanged. After analyzing the prediction label and the prediction probability, we find an interesting phenomenon that $\cos\theta$ affects the prediction probability while does not affect the prediction label. This seems to imply that if we pay the same attention to text and image, the QIMF strategy has no significant effect on the accuracy. When $\alpha^2 = 0.4\, and\, \beta^2 = 0.6$ or $\alpha^2 = 0.3\, and\, \beta^2 = 0.7$, the accuracy increases until $\cos\theta = 1$. Further, for both GI and Flickr datasets, we can observe that our QMSA framework reaches the best performances, when $\cos\theta = 0.3$ and $\cos\theta = 0.15$, respectively.

In Figure 7, we notice the similar evidence that the accuracy is the highest when $\alpha^2 =$

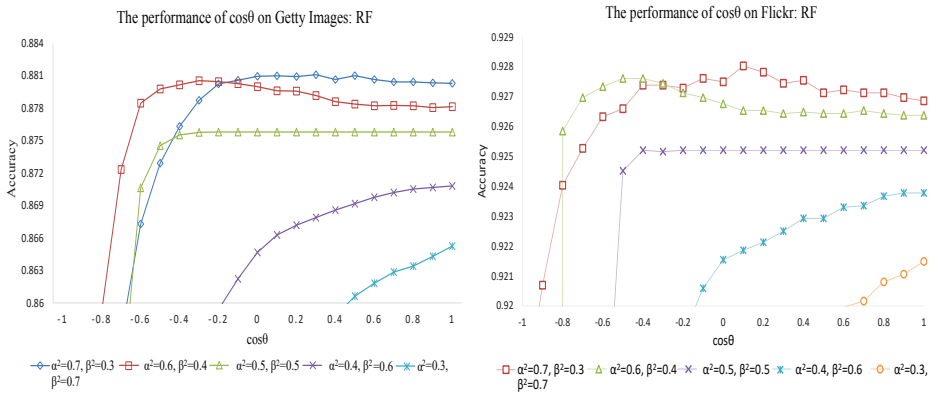


Figure 6: The effect of $\cos\theta$ on Getty Images and Flickr datasets using RF classifiers.

$0.8\,and\,\beta^2 = 0.2$ or $\alpha^2 = 0.7\,and\,\beta^2 = 0.3$ on the two datasets. When $\alpha^2 = 0.3\,and\,\beta^2 = 0.7$, the accuracy is the lowest. We can observe that our QMSA framework reaches the best performance, when $\cos\theta = -0.6$ and $\cos\theta = -0.9$, respectively. For the Getty Images dataset, when $\alpha^2 = 0.8\,and\,\beta^2 = 0.2$ and $\alpha^2 = 0.7\,and\,\beta^2 = 0.3$, the accuracy when $\cos\theta = -0.6$ is higher than the accuracy when $\cos\theta = 0$, by about 5%. When $\alpha^2 = 0.5\,and\,\beta^2 = 0.5$, the accuracy increases sharply until $\cos\theta = -0.8$ and $\cos\theta = -0.3$, then keeps nearly unchanged. From Figure 6 and Figure 7, we can conclude that the sentiment polarities of images are consistent with the polarities of texts for most multimodal documents. These results have showed that the influence of different $\cos\theta$ on the classification results.

7. Conclusions

Multimodal sentiment analysis is an important but challenging task. In this paper, we propose a Quantum-inspired Multimodal Sentiment Analysis (QMSA) framework, which contains a Quantum-inspired Multimodal Representation (QMR) model and a Quantum Interference inspired Multimodal Decision Fusion (QIMF) strategy. In our framework, both the text and the image are associated to density matrices, which are estimated by a globally convergent algorithm. Furthermore, the complementary decision information is considered through adding an interference term at the decision level. We apply the QMR model to extract both textual and visual features, then use the QIMF strategy to make the final decision about the sentiment category. The experimental results on two large scale datasets, which are crawled from the Getty Images and Flickr photo sharing platform, demonstrate that our proposed framework largely outperforms a number of state-of-art sentiment analysis algorithms.

Acknowledgements. This work is supported in part by the Chinese National Program on Key Basic Research Project (973 Program, grant No. 2014CB744604), Natural Science Foundation of China (grant No. U1636203, 61272265, 61402324), and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721321.

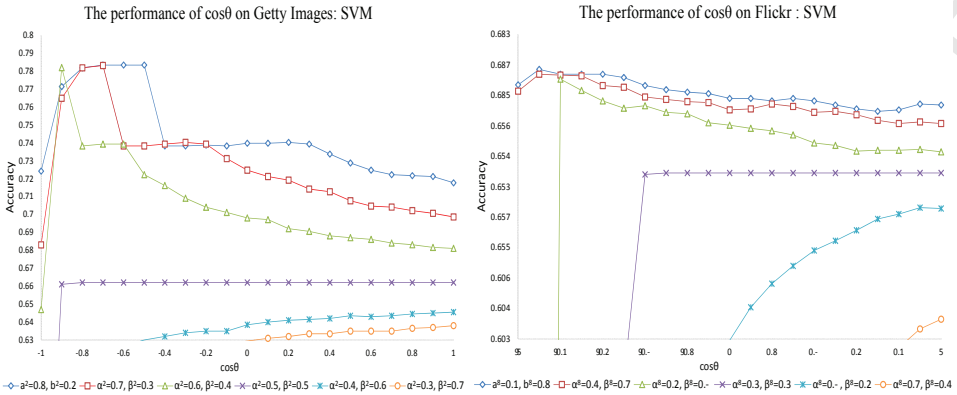


Figure 7: The effect of $\cos\theta$ on Getty Images and Flickr datasets using SVM classifiers.

References

[1] 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.

[2] M. P. Villegas, M. J. Garciarena Ucelay, J. P. Fernández, M. A. Álvarez Carmona, M. L. Errecalde, L. Cagnina, Vector-based word representations for sentiment analysis: a comparative study, in: XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)., 2016.

[3] R. Mihalcea, Multimodal sentiment analysis, in: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, 2012, pp. 1–1.

[4] M. S. Vohra, J. Teraiya, Applications and challenges for sentiment analysis: A survey, in: International Journal of Engineering Research and Technology, Vol. 2, ESRSA Publications, 2013.

[5] C. Baecchi, T. Uricchio, M. Bertini, A. Del Bimbo, A multimodal feature learning approach for sentiment analysis of social network multimedia, Multimedia Tools and Applications 75 (5) (2016) 2507–2525.

[6] 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.

[7] V. Rosas, R. Mihalcea, L.-P. Morency, Multimodal sentiment analysis of spanish online videos, IEEE Intelligent Systems 28 (3) (2013) 38–45.

[8] 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.

[9] M. H. Pereira, F. L. Pádua, A. Pereira, F. Benevenuto, D. H. Dalip, Fusing audio, textual and visual features for sentiment analysis of news videos, arXiv preprint arXiv:1604.02612.

[10] H. Abburi, E. S. A. Akkireddy, S. V. Gangashetty, 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.

[11] J. Wang, Z. Liu, Y. Wu, Learning actionlet ensemble for 3d human action recognition, in: Human Action Recognition with Depth Cameras, Springer, 2014, pp. 11–40.

[12] A. Shahroudy, G. Wang, T.-T. Ng, Multi-modal feature fusion for action recognition in rgb-d sequences, in: Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on, IEEE, 2014, pp. 1–4.

[13] W. Heisenberg, et al., Physics and philosophy, Prometheus Books,, 1999.

[14] P. Suppes, The probabilistic argument for a nonclassical logic of quantum mechanics, in: Studies in the Methodology and Foundations of Science, Springer, 1969, pp. 243–252.

[15] Kolmogorov, Foundations of the theory of probability.

[16] J. V. Neumann, Mathematical foundations of quantum mechanics, no. 2, Princeton university press, 1955.

[17] P. D. Bruza, Z. Wang, J. R. Busemeyer, Quantum cognition: a new theoretical approach to psychology, Trends in cognitive sciences 19 (7) (2015) 383–393.

[18] M. Melucci, Can information retrieval systems be improved using quantum probability?, in: Conference on the Theory of Information Retrieval, Springer, 2011, pp. 139–150.
[19] D. Aerts, J. A. Arguëlles, L. Beltran, L. Beltran, M. S. de Bianchi, S. Sozzo, T. Veloz, Context and interference effects in the combinations of natural concepts, arXiv preprint arXiv:1612.06038.
[20] D. Aerts, P. Bruza, Y. Hou, J. Jose, M. Melucci, J.-Y. Nie, D. Song, Quantum theory-inspired search, Procedia Computer Science 7 (2011) 278–280.
[21] J. R. Busemeyer, Z. Wang, What is quantum cognition, and how is it applied to psychology?, Current Directions in Psychological Science 24 (3) (2015) 163–169.
[22] D. Aerts, L. Gabora, S. Sozzo, Concepts and their dynamics: A quantum–theoretic modeling of human thought, Topics in Cognitive Science, 5, pp. 737-772.
[23] J. R. Busemeyer, P. D. Bruza, Quantum models of cognition and decision, Cambridge University Press, 2012.
[24] B. Wang, P. Zhang, J. Li, D. Song, Y. Hou, Z. Shang, Exploration of quantum interference in document relevance judgement discrepancy, Entropy 18 (4) (2016) 144.
[25] M. Melucci, B. Piwowarski, Quantum mechanics and information retrieval: From theory to application, in: Proceedings of the 2013 Conference on the Theory of Information Retrieval, ACM, 2013, p. 1.
[26] M. Melucci, When index term probability violates the classical probability axioms quantum probability can be a necessary theory for information retrieval, arXiv preprint arXiv:1203.2569.
[27] P. Zhang, J. Li, B. Wang, X. Zhao, D. Song, Y. Hou, M. Melucci, A quantum query expansion approach for session search, Entropy 18 (4) (2016) 146.
[28] B. Piwowarski, I. Frommholz, M. Lalmas, K. Van Rijsbergen, What can quantum theory bring to information retrieval, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 2010, pp. 59–68.
[29] D. Song, M. Lalmas, K. Van Rijsbergen, I. Frommholz, B. Piwowarski, J. Wang, P. Zhang, G. Zuccon, P. Bruza, S. Arafat, et al., How quantum theory is developing the field of information retrieval., in: AAAI Fall Symposium: Quantum Informatics for Cognitive, Social, and Semantic Processes, 2010.
[30] V. Hatzivassiloglou, K. R. McKeown, Predicting the semantic orientation of adjectives, in: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1997, pp. 174–181.
[31] P. D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[32] J. Wiebe, Learning subjective adjectives from corpora, in: AAAI/IAAI, 2000, pp. 735–740.
[33] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 168–177.
[34] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining., in: LREC, Vol. 10, 2010, pp. 2200–2204.
[35] T. I. Jain, D. Nemade, Recognizing contextual polarity in phrase-level sentiment analysis, International Journal of Computer Applications IJCA 7 (5) (2010) 5–11.
[36] P. Stone, D. C. Dunphy, M. S. Smith, D. Ogilvie, The general inquirer: A computer approach to content analysis, Journal of Regional Science 8 (1) (1968) 113–116.
[37] C. Musto, G. Semeraro, M. Polignano, A comparison of lexicon-based approaches for sentiment analysis of microblog posts, Information Filtering and Retrieval 59.
[38] A. Moreno-Ortiz, C. P. Hernández, Lexicon-based sentiment analysis of twitter messages in spanish, Procesamiento del lenguaje natural 50 (2013) 93–100.
[39] S. Trinh, L. Nguyen, M. Vo, P. Do, Lexicon-based sentiment analysis of facebook comments in vietnamese language, in: Recent Developments in Intelligent Information and Database Systems, Springer, 2016, pp. 263–276.
[40] A. Cui, H. Zhang, Y. Liu, M. Zhang, S. Ma, Lexicon-based sentiment analysis on topical chinese microblog messages, in: Semantic Web and Web Science, Springer, 2013, pp. 333–344.
[41] H. Saif, Y. He, M. Fernandez, H. Alani, Contextual semantics for sentiment analysis of twitter, Information Processing & Management 52 (1) (2016) 5–19.
[42] D. R. Recupero, V. Presutti, S. Consoli, A. Gangemi, A. G. Nuzzolese, Sentilo: frame-based sentiment analysis, Cognitive Computation 7 (2) (2015) 211–225.
[43] A. Gangemi, V. Presutti, D. R. Recupero, Frame-based detection of opinion holders and topics: a model and a tool, IEEE Computational Intelligence Magazine 9 (1) (2014) 20–30.
[44] D. R. Recupero, S. Consoli, A. Gangemi, A. G. Nuzzolese, D. Spampinato, A semantic web based core engine to efficiently perform sentiment analysis, in: European Semantic Web Conference, Springer, 2014, pp. 245–248.
[45] V. Presutti, D. Reforgiato, A. Gangemi, A. G. Nuzzolese, S. Consoli, D. Spampinato, Sentilo: Semantic web-based sentiment analysis.
[46] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in:

Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, 2002, pp. 79–86.

[47] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining., in: LREc, Vol. 10, 2010, pp. 1320–1326.

[48] O. Kolchyna, T. T. Souza, P. Treleaven, T. Aste, Twitter sentiment analysis: Lexicon method, machine learning method and their combination, arXiv preprint arXiv:1507.00955.

[49] A. Z. Khan, M. Atique, V. Thakare, Combining lexicon-based and learning-based methods for twitter sentiment analysis, International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE) (2015) 89.

[50] S. Siersdorfer, E. Minack, F. Deng, J. Hare, Analyzing and predicting sentiment of images on the social web, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 715–718.

[51] M. Zubair Asghar, S. Ahmad, A. Marwat, F. Masud Kundi, Sentiment analysis on youtube: A brief survey, arXiv preprint arXiv:1511.09142.

[52] M. Sikandar, A survey for multimodal sentiment analysis methods, International Journal Computer Technology & Applications 5 (4) (2014) 1470–1476.

[53] S. Poria, A. Hussain, E. Cambria, Beyond text based sentiment analysis: Towards multi-modal systems, University of Stirling, Stirling FK9 4LA, UK, Tech. Rep.

[54] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of EMNLP, 2015, pp. 2539–2544.

[55] 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.

[56] D. Maynard, D. Dupplaw, J. Hare, Multimodal sentiment analysis of social media.

[57] Y. Yu, H. Lin, J. Meng, Z. Zhao, Visual and textual sentiment analysis of a microblog using deep convolutional neural networks, Algorithms 9 (2) (2016) 41.

[58] F. Chen, Y. Gao, D. Cao, R. Ji, Multimodal hypergraph learning for microblog sentiment prediction, in: 2015 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2015, pp. 1–6.

[59] R. H. Landau, Quantum mechanics II: a second course in quantum theory, John Wiley & Sons, 2008.

[60] B. Simmons, Operator methods in quantum mechanics.

[61] 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.

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

[63] 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.

[64] D. S. Goncalves, M. A. Gomes-Ruggiero, C. Lavor, Global convergence of diluted iterations in maximum-likelihood quantum tomography, arXiv preprint arXiv:1306.3057.

[65] C. Zhai, Statistical language models for information retrieval, Synthesis Lectures on Human Language Technologies 1 (1) (2008) 1–141.

[66] R. P. Feynman, R. B. Leighton, M. Sands, The Feynman Lectures on Physics, Desktop Edition Volume I, Vol. 1, Basic books, 2013.

[67] D. Dieks, The formalism of quantum theory: an objective description of reality?, Annalen der Physik 500 (3) (1988) 174–190.

[68] Q. You, L. Cao, H. Jin, J. Luo, Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks, in: Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016, pp. 1008–1017.

[69] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50, `http://is.muni.cz/publication/884893/en`.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[71] M. Neethu, R. Rajasree, Sentiment analysis in twitter using machine learning techniques, in: Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, IEEE, 2013, pp. 1–5.

[72] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 524–531.

[73] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multi-

task learning, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 160–167.

[74] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[75] J. Kittler, M. Hatef, R. P. Duin, J. Matas, On combining classifiers, IEEE transactions on pattern analysis and machine intelligence 20 (3) (1998) 226–239.

[76] D. Pan, P. Zhang, J. Li, D. Song, J.-R. Wen, Y. Hou, B. Hu, Y. Jia, A. De Roeck, Using dempster-shafer's evidence theory for query expansion based on freebase knowledge, in: Asia Information Retrieval Symposium, Springer, 2013, pp. 121–132.

[77] N. Srivastava, R. R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: Advances in neural information processing systems, 2012, pp. 2222–2230.

[78] J. Urban, J. M. Jose, C. J. van Rijsbergen, An adaptive technique for content-based image retrieval, Multimedia Tools and Applications 31 (1) (2006) 1–28. doi:10.1007/s11042-006-0035-1.
URL https://doi.org/10.1007/s11042-006-0035-1

(a) girl white dress holding out hands with floating white paper crane

(b) woman admiring art gallery

(c) khon pantomime performances action thailand kind thai drama hanuman monkey performing action with beautiful sida wat phra khao temple ayutthaya province

(d) two rockers attracted to each other giving sideways glances

(e) climber offers teammate helping hand mtns

(f) rusty bollard with chain footpath

(g) child holding gun mid-section bw

(h) mixed race businesswoman kneeling railroad tracks

(i) suicide letter parchment paper with blood drops and quill

(j) professional containing contagious patient hospital during pandemic

Figure 8: Several examples of Getty Images dataset using QMSA framework. Multimodal documents of the first line are positive examples; multimodal documents of the second line are negative examples.

(a) dltbco greyhound

(b) christmas perfume ad nina ricci lairs du temps fragrance

(c) two models next lockheed electra delta air lines

(d) blackpool air port thornycroft crash tender dennis body work

(e) working pitcairn terminal

(f) My Bloody Valentine

(g) Injured Negro woman assisted by Negro police officer

(h) worry dsc

(i) scary night done

(j) domestic violence month press conference

Figure 9: Several examples of Flickr dataset using QMSA framework. Multimodal documents of the first line are positive examples; multimodal documents of the second line are negative examples.