Neurocomputing 445 (2021) 276–286

Contents lists available at ScienceDirect

Neurocomputing

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

Quantum probability-inspired graph neural network for document representation and classification

Peng Yan [a,b], Linjing Li [a,b,c,*], Miaotianzi Jin [c], Daniel Zeng [a,b,c]

[a] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[c] Shenzhen Artificial Intelligence and Data Science Institute, Shenzhen, China

ARTICLE INFO

Article history:
Received 7 September 2020
Revised 13 January 2021
Accepted 22 February 2021
Available online 9 March 2021

Keywords:
Natural language processing
Document representation
Document classification
Graph neural network
Quantum probability

ABSTRACT

Recent studies have found that text can be represented in Hilbert space through a neural network driven by quantum probability, which provides a unified representation of texts with different granularities without losing the performance of downstream tasks. However, these quantum probability-inspired methods only focus on intra-document semantics and lack modeling global structural information. In this paper, we explore the potential of combining quantum probability with graph neural network, and propose a quantum probability-inspired graph neural network model to capture global structural information of interaction between documents for document representation and classification. We build a document interaction graph for a given corpus based on document word relation and frequency information, then learn a graph neural network driven by quantum probability on the defined graph. First, the proposed model represents each document node in the graph as a superposition state in a Hilbert space. Then the proposed model further computes density matrix representations for nodes to encode document interaction as mixed states. Finally, the model computes classification probability by performing quantum measurement on the mixed states. Experiments on four document classification benchmarks show that the proposed model outperforms a variety of classical neural network models and the previous quantum probability-inspired model with much smaller parameter size. Extended analyses also demonstrate the robustness of the proposed model with limited training data and its ability to learn semantically distinguishable document representation.

1. Introduction

Text representation converts the discrete space of natural language into continuous semantic space for further processing, which plays a fundamental role in natural language processing (NLP). Over the last few years, neural network models for text representation have been widely adopted in a variety of areas of NLP, e.g., text classification [1,2], machine translation [3,4], question answering [5], sentiment analysis [6,7], and other applications. Most of the existing neural network language models rely on word representations like Word2Vec [8] or GloVe [9] to capture word semantics, and focus on developing sophisticated compositional architectures of word representations to model document semantics, such as convolutional neural networks (CNNs) [1], recurrent neural networks (RNNs) like long short-term memory (LSTM) [10], and Transformer [11]. Recently, Hilbert Semantic Space for text representation is proposed to model human language with a well-designed mathematical framework of quantum probability [12,13]. The new framework unifies different linguistic units (e.g. sememe, word, sentence) in a single Semantic Hilbert Space and demonstrates interpretations to explicit physical meanings without losing the performance of downstream tasks, including text classification and matching.

However, this quantum probability-inspired framework is implemented as end-to-end neural network architecture and follows the practice of neural network language models to build compositional architectures on word representations, thus it may face three challenges that need to be addressed:

- First, word representations cannot capture semantic information effectively in some practical applications due to data sparsity problem. For example, some datasets, such as medical records and industry-specific news, are often limited in size

* Corresponding author at: Institute of Automation, Chinese Academy of Sciences, Beijing, China.
E-mail addresses: yanpeng2017@ia.ac.cn (P. Yan), linjing.li@ia.ac.cn (L. Li), jin@saidi.org.cn (M. Jin), dajun.zeng@ia.ac.cn (D. Zeng).

https://doi.org/10.1016/j.neucom.2021.02.060
0925-2312/

and containing a large number of out-of-vocabulary (OOV) words such as professional nouns, as illustrated in Table 1. Therefore, representations of OOV words cannot be initialized with pre-trained word representations and have to be given random values instead. Thus, the semantics of these words totally relies on fine-tuning on the downstream tasks. But the fine-tuning process may be ineffective due to the sparsity of words in the dataset, where the average of occurrences of OOV words is often extremely low, as shown in Table 1.

- Second, compositional architectures on word representations in neural network models only focus on capturing intra-document semantics, including sequential syntactic and local semantic information [14], but ignore global structural information of document semantic interaction, which can be called inter-document semantics, including long-distance and non-consecutive semantic information in a corpus.
- Third, both word representations and sophisticated compositional architectures require a substantial scale of parameters and massive computation. Particularly, word representations often account for a dominating part of the model parameter size [15]. Training neural network models with the oversized word representations on the limited-scale dataset not only wastes computing resources and power but also causes overfitting problems.

To address these challenges, we explore the potential of combining quantum probability with graph neural networks, which have been shown effectiveness in a wide range of applications because they pay more attention to structural information than normal neural networks [16,17]. The significance of modeling human language with quantum probability is confirmed by extensive literature, which suggests that quantum-like phenomena exist in human cognition [18], decision-making [19], and natural language [20,21], but the quantum probability framework needs to be implemented as a more effective neural network architecture, and its combination with graph neural networks have not yet been explored.

In this paper, we propose a quantum probability-inspired graph neural network, named *Doc2Ket*, to capture global structural information of document semantic interaction for document representation and classification. Specifically, we construct a graph to describe the semantic interaction of documents from an entire corpus by considering documents as nodes and computing edge weights based on document word relation and frequency information. And then we apply a novel graph neural network model driven by quantum probability to directly learn document representation on the defined document interaction graph, instead of modeling documents as compositional semantics of word representations. First, the proposed model considers each document node as a particle, different semantic meanings (or latent concepts) as different basis states to define a semantic representation space. Thus, the model can represent each document node as a superposition state in the space, denoted as a ket in Dirac notation in quantum mechanics. Then the model treats the contextuality of each document node and neighboring document nodes in the interaction graph as mixed systems and computes density matrix representations of mixed states to capture inter-document semantic information. Finally, the model performs a set of measurements on the mixed states to generate measurement probability for each classification category, which can be seen as classification probability.

We conduct extensive experiments on four benchmark datasets of the document classification task. The experiments demonstrate that the proposed *Doc2Ket* model can not only outperform a variety of classical neural network models but also achieve better performance in comparison with the previous state-of-the-art quantum probability-inspired model. In addition, we also conduct extended analyses to demonstrate its robustness with limited training data, efficiency in parameter size, and the ability to learn semantically distinguishable document representation.

Table 1
An illustration of the out-of-vocabulary (OOV) problem in the datasets of four text classification benchmarks, which will be introduced in detail in the experiment section. In all four datasets, the OOV words ratios are quite high, where the default vocabulary is set as the union of the vocabularies of GloVe.6B and GloVe.840B embedding [9] and K stands for thousands. At the same time, the average number of occurrences of these OOV words in the whole dataset is extremely low (even lower than 3 times). Furthermore, some examples show that these OOV words are often professional nouns, which are closely related to the classification results.

Datasets	OOV Words Ratio	Average of Occurrences
R8	3.24 K/23.6 K (13.7%)	2.25
R52	3.59 K/26.3 K (13.7%)	2.29
Ohsumed	4.95 K/31.5 K (15.7%)	2.27
20NG	46.7 K/123 K (38.0%)	2.89
Datasets		Some Examples of OOV Words
R8		ameritrust, carloadings, intermedics
R52		naturalite, equiticorp, endotronics
Ohsumed		dopexamine, insulinopenic, tumourlets
20NG		telecomputing, stylewriter, colostate, symbiotics

2. Related work

2.1. Quantum probability-inspired methods for NLP

In this paper, the proposed model is inspired by the recent multidisciplinary research across quantum probability and NLP. Quantum probability refers to the mathematical foundation of quantum mechanics, which can not only explain non-classical behaviors of microscopic particle in physics, but also be in principle applicable in macro-world problems that need to formalize uncertainty. Indeed, researchers have been exploring applications of quantum probability in psychology [22], cognition and decision making [23–25], IR [26,27] and NLP [28,29,12].

In information retrieval (IR), the first attempt to adopt quantum probability was to reforge various IR formal models in the quantum-theoretic framework and provide a sound basis for developing new models to address IR problems, such as pseudo-relevance feedback and ostensive retrieval [26]. Later, Piwowarski et al. proposed a quantum-like IR model to represent queries as density matrices and documents as subspaces [30]. Then, Quantum Probability Ranking Principle (QPRP) for re-ranking top retrieved documents was developed to implicitly capture inter-dependencies between documents, based on the analogies between the inter-document dependency and quantum interference phenomena [31,32]. Recently, Bengio et al. proposed a principled Quantum Language Model (QLM), which generalizes the traditional language model with quantum probability [27]. QLM represents a query or document as a density matrix to compute ranking metrics and estimates density matrix based on Maximal Likelihood Estimation (MLE), providing a significant improvement on realistic ad hoc retrieval tasks. Lately, Yoshua Bengio et al. further extended QLM to learn concept embeddings for query expansion inspired by quantum entropy minimization [33]. And Li et al. proposed a Session-based Quantum Language Model (SQLM) based on density matrix transformation to capture the dynamic information in multi-query session search task [34].

The successful application of the quantum probability-inspired models in IR has led researchers to further explore their application in NLP. In recent years, quantum probability-inspired models for NLP have been attracting considerable attention with successful applications in various tasks, such as sentiment analysis [35,36], question answering [37], and other applications. Lapata et al. proposed a quantum probability-inspired model [28] by defining a dependency-based Hilbert space and adopting density matrices to encode dependency neighborhood, for modeling context effects in word similarity and association task. But this framework does not have connections with the neural network design. With the popularity of deep learning in NLP, the Neural Network-based Quantum Language Model (NNQLM) [37] has been proposed by integrating quantum probability into an end-to-end neural network structure. NNQLM designs an end-to-end neural network based on the density matrix to model question–answer pairs for QA tasks. Lately, a quantum probability-driven network (QPDN) [12,13] has been further proposed to model different levels of semantic units, including sememes, word, document, and semantic abstraction for natural language understanding.

However, the previous quantum probability-inspired models focus modeling intra-document semantics by compositionality of word embeddings. Different from them, our model focuses on modeling inter-document semantics in the corpus by combining the framework of quantum probability with graph neural network architecture to directly learn document representation for classification. Specifically, as shown in Table 2, the concepts in quantum probability have different roles in modeling semantics between the previous quantum probability-inspired model, QPDN [12] and the proposed model, *Doc2Ket*.

2.2. Graph neural networks

Graph Neural Networks (GNNs) have received massive attention in recent years and have been applied to a wide range of NLP tasks [16,17]. Graph Convolutional Network (GCN) that convolves features of neighbors, is proposed by generalizing CNN on the grid structure to arbitrarily structured graphs [38]. More recently, graph attention network (GAT) is proposed to aggregate neighborhood features with masked self-attentional layers [39]. Compared to them, motivation and network structure of the proposed model are different. In particular, the proposed model aggregates neighborhood information to update each node representation by computing density matrix of mixed system inspired by quantum probability. And the proposed model implements aggregation operation as mixture, downstream classification as measurement in a single unified framework of quantum probability with explicit physical meaning.

Moreover, some studies have combined quantum computing with neural networks [40] and graph neural networks [41,42], and developed a few models with applications to different tasks, including network analysis [43] and image recognition [44]. Different from these works, the proposed quantum probability-inspired graph neural network is developed from the quantum probability-driven network (QPDN) for natural language modeling [12,13], hence it focuses on the applications to text representation and classification.

Finally, GNNs have also been very popular in NLP tasks, including word representations [17], machine reading comprehension [45], document summarization [46], and text classification [47]. Compared to the previous works on text classification task, besides differences in network architecture, the proposed model constructs a novel document interaction graph based on TF-IDF without requiring extra inter-document relations such as citation relation, and treats a document as a node instead of a graph of word nodes.

Table 2
Different roles of concepts between QPDN and *Doc2Ket*.

Concepts	QPDN	Ours: *Doc2Ket*
Basis states	sememes	semantic basis
Superposition	word	document semantics
Mixed system	document	inter-document semantics
Measurement	abstraction	text classification
Measurement probability	high-level representation	classification probability

3. Methodology

3.1. Document interaction graph construction

Knowledge about inter-document relationships in a given corpus is useful for learning document representation for classification. To describe the interactions between documents in the whole corpus $\mathscr{C} = \{d_1, d_2 \cdots d_{m-1}, d_m\}$, we regard the interaction structure as a fully-connected document graph $G = (V, E)$, where V and E are sets of nodes and edges, respectively. Specifically, each node in the graph is a document in the corpus, and every two nodes have an edge. Every node is also added a self-loop edge to itself. The edge weight between two document nodes is computed based on document word relation and frequency information.

In our approach, we set the edge weight w_{ij} as normalized inter-document semantic proximity between document d_i and document d_j, which is computed as the sum of products of term frequency-inverse document frequency (TF-IDF) between all the co-occurrence vocabulary and documents, with normalization to guarantee $\sum_{j=1}^{m} w_{ij} = 1$:

$$\tilde{w}_{ij} = \sum_{s=1}^{S_{ij}} \mathrm{TF-IDF}_{i,s} * \mathrm{TF-IDF}_{j,s}, \tag{1}$$

$$w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j=1}^{m} \tilde{w}_{ij}}, \tag{2}$$

where S_{ij} is the size of the co-occurrence vocabulary in d_i and d_j, and $\mathrm{TF-IDF}_{i,s}$ is the TF-IDF score between d_i and the s-th word in the co-occurrence vocabulary. We use TF-IDF score to increase the attention and importance of those low-frequency words, which are often professional nouns or out-of-vocabulary words. Therefore, documents with similar semantic content in the corpus can be linked with high weight in the defined graph. Then, we adopt quantum probability-inspired graph neural network to model the structural information of document interaction graph.

3.2. Basic intuitions of quantum probability-inspired graph neural network

Drawing inspiration from the quantum probability [48], which is a sound mathematical framework of quantum mechanics [49], we propose a novel graph neural network model *Doc2Ket* to capture inter-document semantics in the corpus for document representation and classification.

We first introduce the widely used Dirac notation in quantum mechanics, where a column vector φ is denoted as $|\varphi\rangle$, called as a ket. The transpose of $|\varphi\rangle$ (a row vector) is denoted as $\langle\varphi|$, called as a bra. The inner product of two state vectors $|\varphi_1\rangle$ and $|\varphi_2\rangle$ is denoted as $\langle\varphi_1|\varphi_2\rangle$, and the outer product of them is denoted as $|\varphi_1\rangle\langle\varphi_2|$.

The basic intuitions of the proposed model are illustrated in Fig. 1.

- First, inspired by the framework of quantum probability, we consider each document node as a particle, different semantic meanings (or latent concepts) as different basis states. Superposition state $|\varphi\rangle$ can model uncertainty of a particle's state that is superposed in different basis states $\{|x_i\rangle\}_{i=1}^{n}$. Analogously, a document can contain multiple meanings. Based on such an analogy, to describe each document node superposed in multiple semantic basis vectors (for multiple meanings/concepts), the proposed model directly models each document node as a superposition state: $|\varphi\rangle = \sum_{i=1}^{n}\alpha_i|x_i\rangle$, where probability amplitudes $\{\alpha_i\}_{i=1}^{n}$ are complex-valued scalars satisfying $0 \leqslant |\alpha_i|^2 \leqslant 1$ and $\sum_{i=1}^{n}|\alpha_i|^2 = 1$. Superposition state can be considered as a ray in a Hilbert space, which is a complete vector space defined on the complex-valued scalar field endowed with an inner product.
- Second, we consider the contextuality of the document node and its neighborhood in the document interaction graph as mixed systems. In quantum probability, a density matrix can model the state of a mixed system of multiple particles $\left\{|\varphi_j\rangle\right\}_{j=1}^{m}$ and is computed as: $\rho = \sum_{j=1}^{m}\omega_j|\varphi_j\rangle\langle\varphi_j|$, where ω_j is the proportion of participants. Hence, we use such a density matrix to represent the mixed state for each document node and to absorb inter-document semantic information from its neighborhood in the graph, instead of using convolution or attention operations in Graph Convolutional Network (GCN) [38] or Graph Attention Network (GAT) [39].
- Finally, we follow the practice of quantum probability to extract the probabilistic properties of the system by quantum measurement. The Gleason's Theorem [50] can be used to calculate the measurement probability as $p_v(\rho) = \langle v|\rho|v\rangle = tr(\rho|v\rangle\langle v|)$, where $|v\rangle$ is a measurement state. The proposed model performs measurements of different measurement states on the density matrix representation of the mixed state for each document node, to compute the classification probabilities for different categories, where each measurement state corresponds to one category.

The whole model based on the above intuitions is implemented as an efficient and effective graph neural network architecture to work in an end-to-end learning mode, as illustrated in Fig. 2. Components of the proposed neural network architecture, including semantic representation, semantic mixture, and semantic measurement, will be introduced in detail as follows.

3.3. Semantic representation

Given the document interaction graph of a corpus, the proposed model follows the framework of quantum probability to define a semantic representation space H as a n-dimensional real-valued vector space. We assume H is spanned by a set of mutually orthogonal unit-length vectors $\{|e_i\rangle\}_{i=1}^{n}$ as semantic basis states, which are the minimum semantic units of text meanings. Although quantum probability is established on the complex-valued scalar field, we restrict vector space to the real-valued scalar field for simplicity and being in line with the previous quantum probability-inspired models [37,51].

To formulate the ambiguous and uncertain semantic composition of a document node, we employ the concept of superposition state. As shown in Fig. 2 Left part, each document node d is modeled as a superposition state over all semantic basis states $\{|e_i\rangle\}_{i=1}^{n}$ in the semantic representation space H:

$$|d\rangle = \sum_{i=1}^{n}\alpha_i|e_i\rangle, \tag{3}$$

where $\{\alpha_i\}_{i=1}^{n}$ are trainable real-valued scalars to learn the most suitable amplitudes automatically by the training data of a specific task. In practical implementation, we ignore the constraints in superposition definition to keep them easier for optimization.

Taking emotion classification as an example, we can define four basic emotions including happiness $|e_h\rangle$, anger $|e_a\rangle$, sorrow $|e_s\rangle$, and fear $|e_f\rangle$ as the semantic basis states, then the semantic representation space H is spanned by these four vectors. The semantics (i.e., emotion) of a document $|d\rangle$ under this task can be represented as a superposition state over the four emotional basis states:

$$|d\rangle = \alpha_h|e_h\rangle + \alpha_a|e_a\rangle + \alpha_s|e_s\rangle + \alpha_f|e_f\rangle. \tag{4}$$

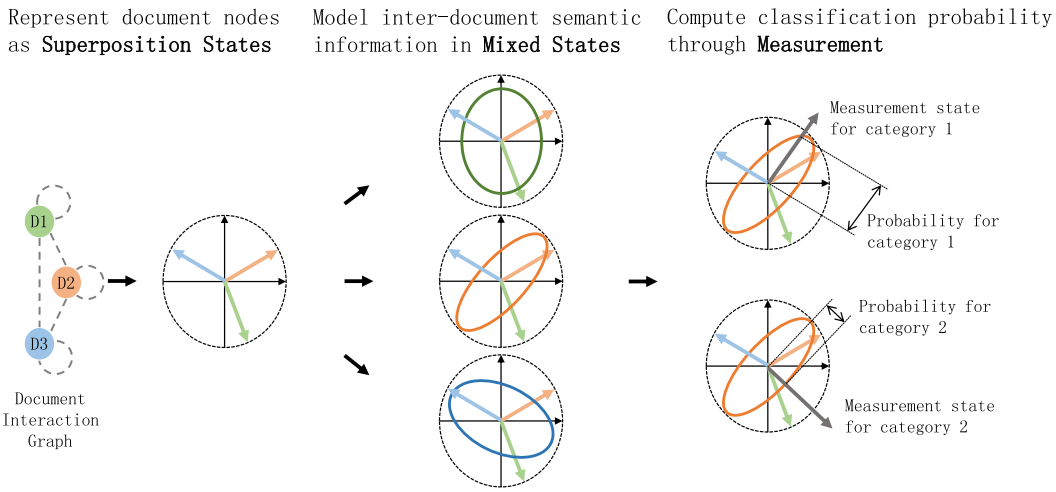


Fig. 1. Basic intuitions of the proposed *Doc2Ket* model. Different colors mean different document nodes modeled as quantum particles (only three examples are shown). The arrows in different colors denote superposition states for document nodes defined in two-dimensional semantic representation space. The ellipses in solid line refer to the quantum probability distributions defined by density matrix representations of mixed systems for documents. The measurements states for different classification categories are in grey lines, the length of which from the origin of coordinates to the intersection with the ellipse corresponds to classification probability.

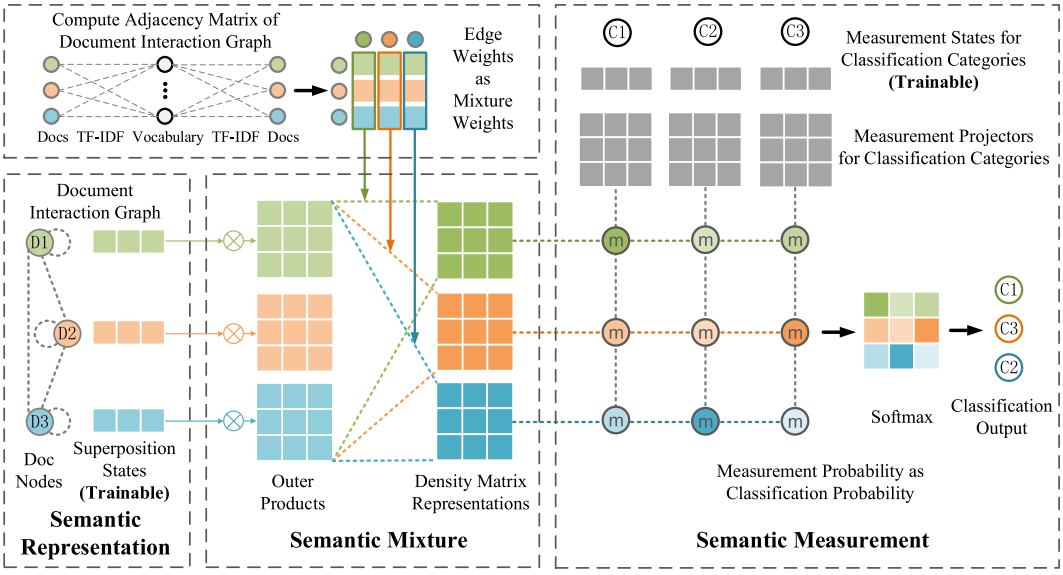


Fig. 2. The overall architecture of the proposed *Doc2Ket* model. The green, orange, and blue colors correspond to three different documents. First, each document node is represented as a trainable superposition state, called ket representation $|d\rangle$ (Left). Second, density matrix representation for each document node is computed as the weighted average of outer products for all document's ket representations (Middle), where mixture weights are edge weights in the document interaction graph. Finally, the model performs a set of measurements for classification categories, illustrated as m, on the density matrix representations of document nodes to generate classification probability (Right).

3.4. Semantic mixture

After modeling document nodes as superposition states, we leverage the concept of mixed state to further capture inter-document semantics in the document interaction graph, as shown in Fig. 2 Middle part. For each document node d_i, we treat the contextuality of document node d_i in the context of its neighborhood as a mixed system, and model the mixed state of the mixed system for d_i as a density matrix ρ_i:

$$\rho_i = \sum_{j=1}^{m} w_{ij} |d_j\rangle\langle d_j|, \tag{5}$$

where $|d_j\rangle$ is superposition state representation for document node d_j in the neighborhood of d_i, and we use the edge weight w_{ij} between d_i and d_j as the mixture weight, which can be understood as the semantic connection between two documents.

The density matrix representation ρ_i for document node d_i is a non-classical probability distribution over semantic basis states, which carries rich semantic information. In particular, the off-diagonal elements provide the potential to describe the correlations and interactions of semantic basis states, giving birth to interference effects of meanings in given states.

3.5. Semantic measurement

After computing density matrices for all document nodes in the graph, we implement text classification by performing a set of measurements on the density matrix representations, as shown in Fig. 2 Right part. Specifically, we first define a set of trainable measurement states $\{|v_j\rangle\}_{j=1}^{c}$ for classification categories, where c is the number of classification categories and each measurement state corresponds to one category. It is worth mentioning that we ignore the orthogonality constraints of the measurement states to keep them trainable, so the most suitable measurement state can be learned automatically for each category. Then the projector for each classification category can be computed as $P_j = |v_j\rangle\langle v_j|$. According to the Gleason's Theorem [50], the measurement probability of performing each measurement projector P_j onto each document node's density matrix ρ_i is computed as:

$$p_{|v_j\rangle}(\rho_i) = \langle v_j|\rho_i|v_j\rangle = tr(\rho_i|v_j\rangle\langle v_j|). \tag{6}$$

After feeding m document nodes in the corpus into the measurement layer of category size c, we can obtain a $m \times c$ matrix of the measurement probabilities $M = \{M_{ij}\}_{i=1,j=1}^{m,c} = \left\{p_{|v_j\rangle}(\rho_i)\right\}_{i=1,j=1}^{m,c}$, as shown in the Fig. 2. Then the measurement probabilities are fed into a softmax classifier to get the final classification probability distribution. The probability of document d_i belonging to the corresponding category of measurement state $|v_j\rangle$ is computed as:

$$Y_{i,j} = \frac{exp(M_{i,j})}{\sum_k exp(M_{i,k})} = \frac{exp\left(p_{|v_j\rangle}(\rho_i)\right)}{\sum_k exp\left(p_{|v_k\rangle}(\rho_i)\right)}. \tag{7}$$

3.6. Training objective

The loss function for training the proposed model consists of two parts. Firstly, the classification loss L_C is designed as a cross-entropy loss between ground truth labels and predicted labels of training set:

$$L_C = -\sum_{i \in Y_T} \sum_{j=1}^{c} \tilde{Y}_{i,j} \ln Y_{i,j}, \tag{8}$$

where Y_T is the set of training document indices, c is the number of categories, and $\tilde{Y}$ is the label indicator matrix, consisting of all one-hot vectors of ground truth labels.

Secondly, due to the fact that we loose the orthogonality constraints of the trainable measurement states $\{|v_j\rangle\}_{j=1}^{c}$ in semantic measurement, we add an orthogonality loss L_O, designed as the average of absolute values of inner products between all two different measurement states, to encourage measurement states to be orthogonal to each other:

$$L_O = \frac{1}{c(c-1)} \sum_{i \neq j} |\langle v_i | v_j \rangle|. \tag{9}$$

In summary, the ultimate loss function L is formulated as:

$$\begin{aligned} L &= L_C + \lambda L_O \\ &= -\sum_{i \in Y_T} \sum_{j=1}^{c} \tilde{Y}_{i,j} \ln Y_{i,j} + \lambda \frac{1}{c(c-1)} \sum_{i \neq j} |\langle v_i | v_j \rangle|, \end{aligned} \tag{10}$$

where λ is the orthogonality parameter for the balance between L_C and L_O.

With the training loss function, we can use the Adam [52], a variant of the back-propagation algorithm, to train the entire model to obtain the values of the trainable parameters, including the superposition states $\{|d_j\rangle\}_{j=1}^{m}$ for all documents and the measurement states $\{|v_j\rangle\}_{j=1}^{c}$ for classification categories.

4. Experiment setup

In this section, we introduce the experiment settings in detail.

4.1. Datasets

We adopt four widely used benchmark datasets including R8, R52, Ohsumed, and 20-Newsgroups (20NG). The statistics of these datasets are summarized in Table 3.

- **R8 and R52:** R8 and R52 are two subsets of the Reuters-21578 collection. R8 includes 7,674 documents for 8 categories, and R52 includes 9,100 documents for 52 categories. [1]
- **Ohsumed:** Ohsumed includes medical abstracts from the MEDLINE database, which is a bibliographic database of medical literature. We use a single-label subset of 7,400 unique abstracts with only one associated category from 23 cardiovascular disease categories. [2]
- **20NG:** 20-Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned evenly across 20 different categories. We use the bydate version of 20-Newsgroups, which contains 18,846 documents, including 11,314 documents for training and 7,532 documents for test.[3]

For all datasets above, we preprocess them by tokenizing text and removing stop words and low-frequency words that appear fewer than 5 times. For fair comparison, we use the same preprocessing methods in all the compared models. As to validation, we randomly select 10% documents from the training set for validation for all four datasets.

4.2. Baselines

To evaluate the performance of *Doc2Ket* model, we conduct a comprehensive comparison with multiple neural network models and the previous quantum probability-inspired model:

- **CNN:** We adopt the CNN model for text classification [1] to perform convolution with different window sizes and max-pooling on word embedding matrix.
- **BiLSTM:** We use two layer bi-directional LSTMs [10] with pre-trained word embeddings for text classification.
- **Transformer:** Transformer [11] uses multi-head self-attention with positional embedding to encode the words.
- **FastText:** FastText is an efficient text classification method [2], which uses the average of word embeddings as document embeddings and feeds them into a linear classifier.
- **CapsuleNet:** Capsule network is originally proposed in image classification domain, but it can be also applied to text classification domain [53].
- **PTE:** Predictive text embedding [54] is a semi-supervised representation learning method for text classification, which constructs a heterogeneous network containing words, documents, and labels as nodes to learn the word and document embedding.
- **LEAM:** Label-embedding attentive model, proposed by [55], utilizes label descriptions to learn words and labels representations in the same space for text classification.
- **SWEM:** Simple word embedding model, proposed by [56], learns document embedding by operating pooling strategies on word embeddings.
- **Graph-CNN:** Graph CNN models operate convolutions over word embedding similarity graphs using different filters, including Chebyshev filter (Graph-CNN-C) [57], Spline filters (Graph-CNN-S) [58], and Fourier filters (Graph-CNN-F) [59].
- **Text-GCN:** A graph convolution network model for text classification [60], learns document embedding by building one unified graph containing word nodes and document nodes for the whole corpus.
- **QPDN:** Quantum Probability Driven Network [12] is a previous quantum probability-inspired neural network model that can encode different levels of semantic units in the same semantic Hilbert space for text classification.

4.3. Parameter settings

For the proposed model, we train it with a learning rate as 0.005 for a maximum of 1000 epochs using Adam [52] and early stop training if the validation loss does not decline for 10 consecutive epochs. For the orthogonality parameter λ, we choose the parameter value in the range of 0 to 0.3 with an interval of 0.01 that achieves the best performance on the validation set and set λ as 0.07 for 20NG, 0.12 for Ohsumed, 0.04 for R8, and 0 for R52. For the dimension n of the semantic representation space H, to provide an enough and appropriate representation space to describe semantic complexity of different corpora, we associate it with the number of classification categories c of the given corpus:

$$n = dimension(H) = int(K * c), \tag{11}$$

where K is the dimension parameter and $int(\cdot)$ means the function to round up to an integer. We choose $K = 1$ as default for all datasets to make the dimension of the measurement space spanned by the measurement states $\{|v_j\rangle\}_{j=1}^{c}$ equals the dimension of the semantic representation space H ($n = c$). For all the baseline models with word embedding, we use 300-dimensional GloVe pre-trained embeddings [9] as initialization and fine-tune the embeddings during training.

5. Result and analysis

5.1. Result

The experimental results are shown in Table 4. *Doc2Ket* model outperforms most baseline models on four datasets and it is comparable to the strong Text GCN model, demonstrating the effectiveness of the proposed model inspired by quantum probability. It is noted that results of some models, including BiLSTM, PTE, LEAM, SWEM, Graph-CNN, Text-GCN, are taken from [60]. For more in-

[1] This dataset is available at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[2] This dataset is available athttp://disi.unitn.it/moschitti/corpora.htm

[3] The dataset is available at http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz

Table 3
Summary statistics of the datasets used in our experiment.

Datasets	#Docs	#Training Docs	#Test Docs	#Categories
R8	7,764	5,485	2,189	8
R52	9,120	6,532	2,568	52
Ohsumed	7,400	3,357	4,043	23
20NG	18,846	11,314	7,532	20

Table 4
Accuracy of different models on four datasets in our experiment. All methods are run ten times and we report mean ± standard deviation of the results.

Model	R8	R52	Ohsumed	20NG
Classical neural network models				
CNN	0.9616 ± 0.0045	0.9032 ± 0.0073	0.5766 ± 0.0117	0.7781 ± 0.0085
BiLSTM	0.9631 ± 0.0033	0.9054 ± 0.0091	0.4927 ± 0.0107	0.7318 ± 0.0185
Transformer	0.9614 ± 0.0050	0.9158 ± 0.0072	0.5933 ± 0.0055	0.7527 ± 0.0127
FastText	0.9616 ± 0.0024	0.9235 ± 0.0027	0.5919 ± 0.0021	0.7944 ± 0.0046
Capsule	0.9688 ± 0.0032	0.9102 ± 0.0025	0.5838 ± 0.0075	0.8318 ± 0.0079
PTE	0.9669 ± 0.0013	0.9071 ± 0.0014	0.5358 ± 0.0029	0.7674 ± 0.0029
LEAM	0.9331 ± 0.0024	0.9184 ± 0.0023	0.5858 ± 0.0079	0.8191 ± 0.0024
SWEM	0.9532 ± 0.0026	0.9294 ± 0.0024	0.6312 ± 0.0055	0.8516 ± 0.0029
Graph-CNN-C	0.9699 ± 0.0012	0.9275 ± 0.0022	0.6386 ± 0.0053	0.8142 ± 0.0032
Graph-CNN-S	0.9680 ± 0.0020	0.9274 ± 0.0024	0.6282 ± 0.0037	–
Graph-CNN-F	0.9689 ± 0.0006	0.9320 ± 0.0004	0.6304 ± 0.0077	–
Text GCN	0.9707 ± 0.0010	0.9356 ± 0.0018	**0.6836 ± 0.0056**	**0.8634 ± 0.0009**
Quantum probability-inspired models				
QPDN	0.9588 ± 0.0056	0.9281 ± 0.0048	0.6326 ± 0.0048	0.7776 ± 0.0050
Ours: *Doc2Ket*	**0.9774 ± 0.0005**	**0.9449 ± 0.0009**	0.6725 ± 0.0008	0.8483 ± 0.0009

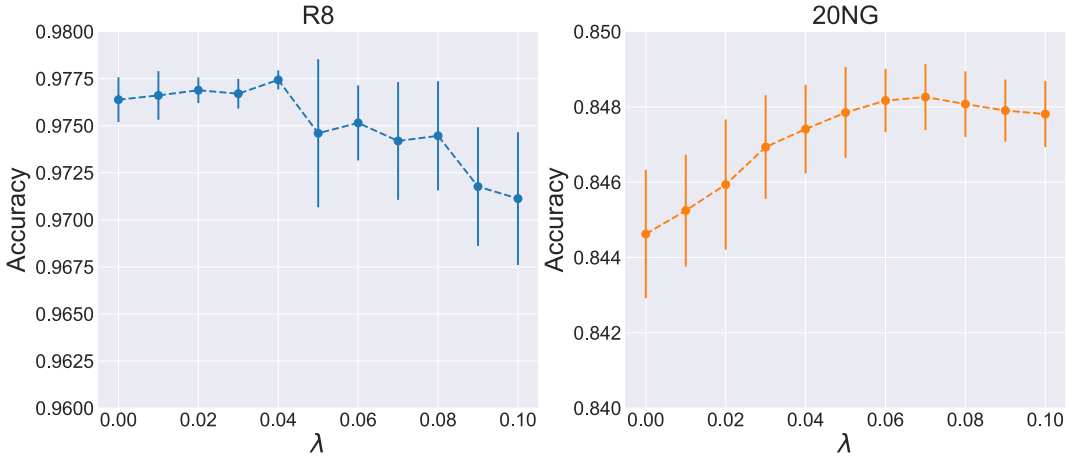


Fig. 3. Sensitivity analysis of orthogonality parameter λ on R8 and 20NG datasets.

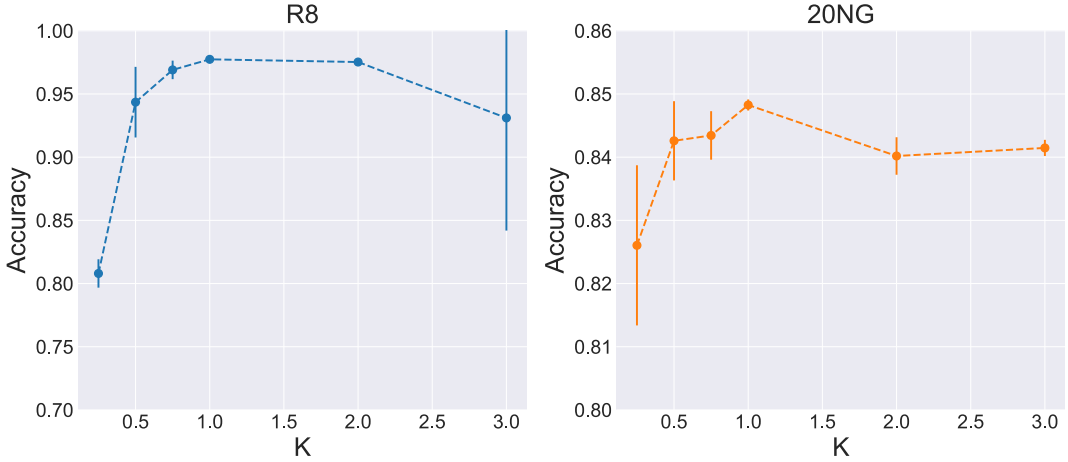


Fig. 4. Sensitivity analysis of dimension parameter K on R8 and 20NG datasets.

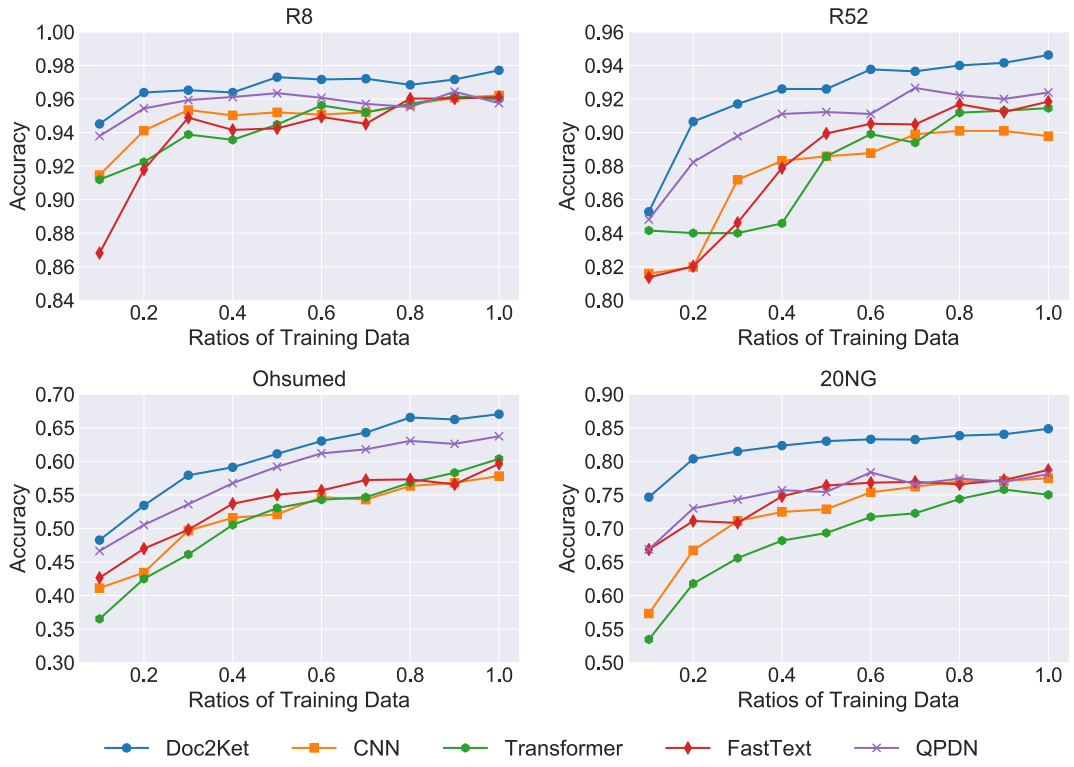


Fig. 5. Analysis on the impact of the training ratio for all datasets. The proposed *Doc2Ket* model performs the best consistently on each dataset and each training ratio.

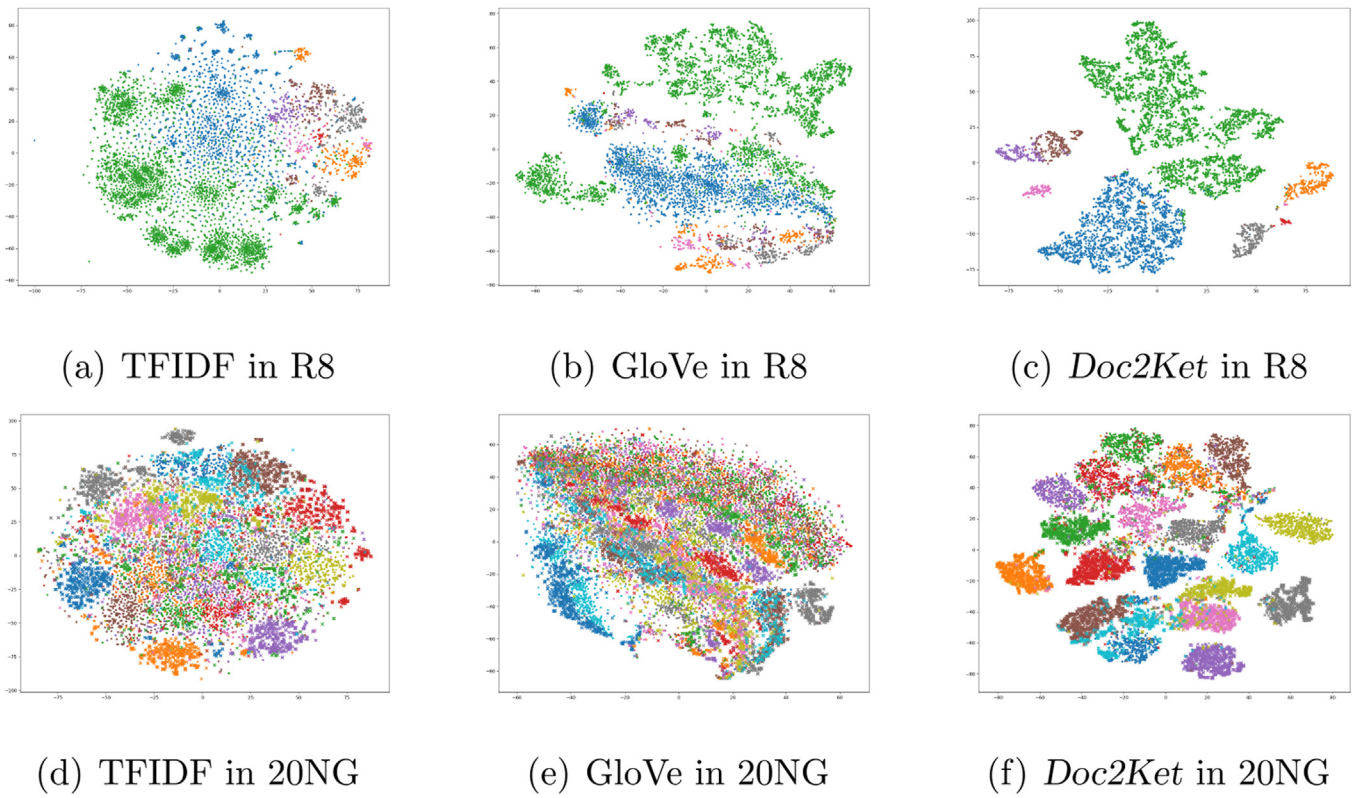


Fig. 6. The visualization result of document representation in R8 and 20NG. Each dot represents a document, and different colors and shapes of dots represent different classification categories.

depth analysis, we note that the performance of *Doc2Ket* model without the use of word embedding can be better than the neural network baselines based on word embedding, including traditional neural networks like CNN, BiLSTM, and advanced neural networks like Transformer and CapsuleNet.

Furthermore, the proposed model also shows competitive performance over the previous quantum probability-inspired method QPDN and a variety of graph convolutional network models, demonstrating the effectiveness of combining quantum probability with graph neural network to model inter-document semantic information in the corpus effectively.

5.2. Sensitivity analysis

Fig. 3 shows test accuracy of *Doc2Ket* with different orthogonality parameter λ in the range of 0 to 0.1 with an interval of 0.01 on R8 and 20NG. We observe that both too small orthogonality parameter λ and too large λ hurt performance. This is consistent with the intuition that too small orthogonality parameter λ can not enforce sufficient orthogonality constraint on measurement states $\{|v_j\rangle\}_{j=1}^{c}$, while too large orthogonality parameter λ may drown out the effect of the classification loss L_C in the total loss L.

Fig. 4 reports test accuracy with different dimension parameter K in the range of $\{0.25, 0.5, 0.75, 1, 2, 3\}$ on R8 and 20NG. We can find that test accuracy first increases as dimension parameter K becomes larger and then decreases after reaching the highest value when $K = 1$. As defined in Eq. (11), dimension parameter K is related to the semantic representation space dimension n. Thus, it is effective to set the dimension of the semantic representation space H to the dimension of the measurement space ($n = c$) by setting K to 1, which is also in line with the normal practice of quantum probability.

5.3. Parameter size

The proposed *Doc2Ket* model shows efficiency in parameter size compared to other neural network models with word embeddings. The trainable parameters of *Doc2Ket* include the superposition states $\{|d_j\rangle\}_{j=1}^{m}$ for all documents, which are $m \times n$ by size, and the measurement states $\{|v_j\rangle\}_{j=1}^{c}$ for classification categories, which are $c \times n$ by size, where m is the document size, n is the dimension of semantic representation space H, and c is the number of categories. In the default configuration, $n = c$, so the parameter size of *Doc2Ket* is $(m + c) \times c$. In contrast, neural network language models need to store the substantial $v \times p$ embedding matrix in GPU memory, where v is vocabulary size and p is dimensionality of the embeddings. For most of limited-scale datasets, document size is smaller than vocabulary size ($m < v$), and the number of categories is usually much smaller that word embedding dimension ($c < p$). Therefore, parameter size of *Doc2Ket* is theoretically smaller than word embedding matrix size of neural network models.

Table 5
Trainable parameter sizes of the proposed model and some neural network baselines (M stands for millions, K stands for thousands).

#Parameter	R8	R52	Ohsumed	20NG
CNN	5.0 M	5.7 M	8.9 M	26.0 M
BiLSTM	4.9 M	5.6 M	8.8 M	26.0 M
FastText	2.9 M	3.4 M	4.9 M	13.5 M
CapsuleNet	10.2 M	29.3 M	20.4 M	36.3 M
QPDN	6.9 M	8.0 M	12.8 M	38.5 M
Ours: *Doc2Ket*	**62.0 K**	**0.6 M**	**180 K**	**0.39 M**

Empirically, we compare trainable parameter sizes of the proposed model with some neural network baselines on four datasets in the experiment. As shown in Table 5, our model uses extremely fewer trainable parameters compared to other baseline models.

5.4. Impact of training ratio

To demonstrate the robustness of the proposed model with limited training data, we further analyze the impact of training ratio by evaluating our model with different ratios of the original training data in the range of 0.1 to 1.0 with an interval of 0.1. We compare the performance of our model with several neural network baselines on all the datasets. As shown in Fig. 5, *Doc2Ket* model can achieve the best performance consistently with different ratios of limited training documents. In particular, *Doc2Ket* achieves a test accuracy of about 0.75 on 20NG with only 10% training documents, which is comparable to the performance of other baselines with even the full training set. This shows the effectiveness and robustness of the proposed model that can utilize limited training data to model inter-document semantics, avoiding the problem of insufficient training of word embedding when data is scarce.

5.5. Visualization

We use t-SNE tool [61] to give an illustrative visualization of the document semantic representation learned by *Doc2Ket* on R8 and 20NG datasets. Specifically, we use the probability distribution for measurement states in the measurement layer as the learned document representation. We also compare the visualization with some document representation baselines. One baseline is TF-IDF that represents documents as term-frequency times inverse document-frequency vectors of the vocabulary. Another is GloVe that represents a document as the average of GloVe embeddings [9] of words in the document. As shown in Fig. 6, we observe that the proposed model *Doc2Ket* can learn more distinguishable and discriminative document representations for different classification categories than the baselines on both R8 and 20NG datasets.

6. Conclusion

In this paper, we further broaden the application of quantum probability in modeling natural language by developing a quantum probability-inspired graph neural network model *Doc2Ket* to leverage inter-document semantics for document representation and classification. We implement the proposed model as an efficient graph neural network in an end-to-end learning mode without the use of word representations and any other sophisticated compositional architectures. The proposed model directly encodes documents into ket representations in semantic representation space, conducts semantic mixture to capture inter-document semantic information among a corpus, and performs semantic measurement for text classification. Experimental results on four benchmark datasets show promising performance of *Doc2Ket* model over multiple strong neural network baselines and the previous quantum probability-inspired model. Besides effectiveness, extended analyses also demonstrate *Doc2Ket*'s robustness to limited training ratio and its efficiency in parameter size.

CRediT authorship contribution statement

Peng Yan: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing - original draft. **Linjing Li:** Supervision, Conceptualization, Methodology, Validation, Writing - review & editing, Resources. **Miaotianzi Jin:** Conceptualization, Validation, Writing - review & editing,

Resources. **Daniel Zeng:** Conceptualization, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0103405, the National Natural Science Foundation of China under Grant 71621002, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27030100, as well as Shenzhen Longhua District Science and Technology Innovation Fund under Grant 10162a20200617b70da63.

References

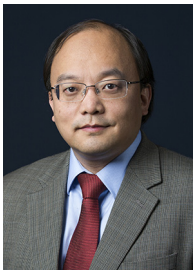
[1] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 2014, pp. 1746–1751..
[2] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp. 427–431..
[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015..
[4] Z. Tan, J. Su, B. Wang, Y. Chen, X. Shi, Lattice-to-sequence attentional neural machine translation models, Neurocomputing 284 (2018) 138–147.
[5] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Advances in neural information processing systems, 2015, pp. 1693–1701..
[6] Z. Cui, X. Shi, Y. Chen, Sentiment analysis via integrating distributed representations of variable-length word sequence, Neurocomputing 187 (2016) 126–132.
[7] J. Wen, G. Zhang, H. Zhang, W. Yin, J. Ma, Speculative text mining for document-level sentiment classification, Neurocomputing..
[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, 2013..
[9] 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), Doha, Qatar, 2014, pp. 1532–1543..
[10] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30, Curran Associates Inc, 2017, pp. 5998–6008..
[12] B. Wang, Q. Li, M. Melucci, D. Song, Semantic hilbert space for text representation learning, The World Wide Web Conference, ACM (2019) 3293–3299.
[13] Q. Li, B. Wang, M. Melucci, CNM: An interpretable complex-valued network for matching, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4139–4148..
[14] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, Q. Yang, Large-scale hierarchical text classification with recursively regularized deep graph-cnn, in: Proceedings of the 2018 World Wide Web Conference, WWW'18, Republic and Canton of Geneva, CHE, 2018, p. 1063–1072..
[15] A. Acharya, R. Goel, A. Metallinou, I. Dhillon, Online embedding compression for text classification using low rank matrix factorization, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6196–6203..
[16] P.W. Battaglia, J.B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V.F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Ç. Gülçehre, H.F. Song, A.J. Ballard, J. Gilmer, G.E. Dahl, A. Vaswani, K.R. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, R. Pascanu, Relational inductive biases, deep learning, and graph networks, CoRR abs/1806.01261.arXiv:1806.01261..
[17] T.T. Tran, M. Miwa, S. Ananiadou, Syntactically-informed word representations from graph neural network, Neurocomputing..
[18] D. Aerts, S. Sozzo, T. Veloz, New fundamental evidence of non-classical structure in the combination of natural concepts, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2058) (2016) 20150095.
[19] C. Moreira, L. Fell, S. Dehdashti, P. Bruza, A. Wichert, Towards a quantum-like cognitive architecture for decision-making, CoRR abs/1905.05176. arXiv:1905.05176..
[20] 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.
[21] D. Aerts, L. Beltran, S. Geriente, S. Sozzo, Quantum-theoretic modeling in computer science: A complex hilbert space model for entangled concepts in corpuses of documents, Int. J. Theor. Phys..
[22] A.Y. Khrennikov, Ubiquitous Quantum Structure: From Psychology to Finance, Springer, 2010.
[23] D. Aerts, L. Gabora, S. Sozzo, Concepts and their dynamics: a quantum-theoretic modeling of human thought, Topics in Cognitive Science 5 (4) (2013) 737–772.
[24] J.R. Busemeyer, P.D. Bruza, Quantum Models of Cognition and Decision, Cambridge University Press, 2013.
[25] P.D. Bruza, Z. Wang, J.R. Busemeyer, Quantum cognition: A new theoretical approach to psychology, Trends in Cognitive Sciences 19 (7)..
[26] C.J.V. Rijsbergen, The geometry of information retrieval, Cambridge University Press, 2004.
[27] 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, SIGIR'13, New York, NY, USA, 2013, p. 653-662..
[28] W. Blacoe, E. Kashefi, M. Lapata, A quantum-theoretic approach to distributional semantics, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 847–857..
[29] I. Basile, F. Tamburini, Towards quantum language models, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, 2017, pp. 1840–1849..
[30] B. Piwowarski, I. Frommholz, M. Lalmas, K. van Rijsbergen, What can quantum theory bring to information retrieval, in: Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26–30, 2010, 2010, pp. 59–68..
[31] X. Zhao, P. Zhang, D. Song, Y. Hou, A novel re-ranking approach inspired by quantum measurement, in: Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18–21, 2011. Proceedings, Vol. 6611 of Lecture Notes in Computer Science, Springer, 2011, pp. 721–724..
[32] G. Zuccon, L. Azzopardi, Using the quantum probability ranking principle to rank interdependent documents, in: Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28–31, 2010. Proceedings, Vol. 5993 of Lecture Notes in Computer Science, Springer, 2010, pp. 357–369..
[33] A. Sordoni, Y. Bengio, J. Nie, Learning concept embeddings for query expansion by quantum entropy minimization, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Québec City, Québec, Canada, 2014, pp. 1586–1592..
[34] 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, Santiago, Chile, August 9–13, 2015, ACM, 2015, pp. 871–874..
[35] Y. Zhang, D. Song, P. Zhang, P. Wang, J. Li, X. Li, B. Wang, A quantum-inspired multimodal sentiment analysis framework, Theoret. Comput. Sci. 752 (2018) 21–40.
[36] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 5436–5442.
[37] P. Zhang, J. Niu, Z. Su, B. Wang, L. Ma, D. Song, End-to-end quantum-like language models with application to question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018..
[38] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings, 2017..
[39] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings, 2018..
[40] C. Shao, Quantum speedup of training radial basis function networks, Quantum Inf. Comput. 19 (7&8) (2019) 609–625.
[41] G. Verdon, T. McCourt, E. Luzhnica, V. Singh, S. Leichenauer, J. Hidary, Quantum graph neural networks, CoRR abs/1909.12264.arXiv:1909.12264..
[42] L. Bai, Y. Jiao, L. Rossi, L. Cui, J. Cheng, E.R. Hancock, Graph convolutional neural networks based on quantum vertex saliency (2019).arXiv:1809.01090..
[43] Z. Zhang, D. Chen, J. Wang, L. Bai, E.R. Hancock, Quantum-based subgraph convolutional neural networks, Pattern Recognit. 88 (2019) 38–49.
[44] Y.C. Li, R.G. Zhou, R.Q. Xu, J. Luo, W.W. Hu, A quantum deep convolutional neural network for image recognition, Quant. Sci. Technol. 5(4)..

[45] B. Zheng, H. Wen, Y. Liang, N. Duan, W. Che, D. Jiang, M. Zhou, T. Liu, Document modeling with graph attention networks for multi-grained machine reading comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, 2020, pp. 6708–6718..
[46] D. Wang, P. Liu, Y. Zheng, X. Qiu, X. Huang, Heterogeneous graph neural networks for extractive document summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, 2020, pp. 6209–6219..
[47] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, Q. Yang, Large-scale hierarchical text classification with recursively regularized deep graph-cnn, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23–27, 2018, 2018, pp. 1063–1072..
[48] J. von Neumann, R. Beyer, Mathematical foundations of quantum mechanics..
[49] J.L. Aronson, The Structure and Interpretation of Quantum Mechanics, Harvard University Press, 1989.
[50] A.M. Gleason, Measures on the closed subspaces of a hilbert space, J. Math. Mech. (1957) 885–893.
[51] T. Wang, Y. Hou, P. Wang, X. Niu, Exploring relevance judgement inspired by quantum weak measurement, AAAI (2018).
[52] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015..
[53] J. Kim, S. Jang, S. Choi, E. Park, Text classification using capsules, arXiv preprint arXiv:1808.03976..
[54] J. Tang, M. Qu, Q. Mei, Pte: Predictive text embedding through large-scale heterogeneous text networks, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'15, New York, NY, USA, 2015, p. 1165-1174..
[55] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, L. Carin, Joint embedding of words and labels for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 2321–2331..
[56] D. Shen, G. Wang, W. Wang, M.R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, L. Carin, Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 440–450..
[57] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain, 2016, pp. 3837–3845..
[58] J. Bruna, W. Zaremba, A. Szlam, Y. LeCun, Spectral networks and locally connected networks on graphs, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings, 2014..
[59] M. Henaff, J. Bruna, Y. LeCun, Deep convolutional networks on graph-structured data, CoRR abs/1506.05163.arXiv:1506.05163..
[60] L. Yao, C. Mao, Y. Luo, Graph convolutional networks for text classification, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019, AAAI Press, 2019, pp. 7370–7377..
[61] G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2605) (2008) 2579–2605.

Peng Yan is currently working towards his PhD degree at the Institute of Automation, Chinese Academy of Sciences, China. He received the B.S. degree in automation from Zhejiang University, Zhejiang, China. His research interests include graph neural network and natural language processing.

Linjing Li is currently an associate professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He received the BE degree in electrical engineering and automation and the ME degree in control theory and control engineering from the Harbin Institute of Technology, Harbin, China, and the PhD degree in computer applied technology from the Graduate University of the Chinese Academy of Sciences, Beijing, China. His research interests include game theory, mechanism design, auction theory, and machine learning. He is a member of the IEEE.

Miaotianzi Jin is currently an assistant professor of Shenzhen Artificial Intelligence and Data Science Institute, Shenzhen, China. He received the BA degree in mathematics and physics from the University of Illinois at Urbana-Champaign and the PhD degree in physics from Northwestern University, Evanston, USA. His research interests include quantum computing, optical computing, and machine learning.



Daniel Zeng is a Research Professor with the Chinese Academy of Sciences, Beijing, China. His research interests include intelligence and security informatics, infectious disease informatics, social computing, recommender systems, software agents, spatial-temporal data analysis, and business analytics. He received the Ph.D. degree in industrial administration from Carnegie Mellon University, Pittsburgh, PA, USA in 1998. He has authored or coauthored one monograph and more than 330 peer-reviewed articles.