IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 December 2024, accepted 17 January 2025, date of publication 24 January 2025, date of current version 30 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3533554

RESEARCH ARTICLE

BERT-Residual Quantum Language Model Inspired by ODE Multi-Step Method

SHAOHUI LIANG[1], YINGKUI WANG[2], AND SHUXIN CHEN[2]

[1]Library, Tianjin Renai College, Tianjin 301636, China
[2]School of Intelligent Computing Engineering, Tianjin Renai College, Tianjin 301636, China

Corresponding author: Shaohui Liang (shaohuiliang@tju.edu.cn)

This work was supported by Tianjin Research Innovation Project for Postgraduate Students under Grant 2021KJ088.

ABSTRACT Quantum-inspired language models model finer-grained semantic interactions in higher-order Hilbert spaces. However, previous methods usually capture semantic features based on context-free word vectors such as Word2Vec and GloVe. Building on natural language encoding, incorporating quantum-inspired density matrix modeling can capture more fine-grained semantic interactions. However, when applied to large pre-trained language models like BERT, using quantum density matrices often leads to issues such as gradient explosion or vanishing. Therefore, how to effectively integrate the quantum-inspired language model and the pre-trained model, and make them function under the fine-tuning paradigm of the pre-trained model has become a key issue for the further development of the quantum-inspired language model. Therefore, in this paper, we propose the BERT-Residual quantum language model inspired by the multi-step method of ordinary differential equations (ODE), using the density matrix to capture the semantic high-order interaction features missing in the BERT modeling process, and obtain the sentence representation, and perform the first step Residuals. Then quantum measurement is performed on the sentence representation, and the second step of residual connection is performed with the BERT layer. This residual connection method based on the multi-step method can more effectively combine the advantages of BERT representation and quantum density matrix representation to enhance representation learning. Experiments show that in text classification benchmarks, our proposed method generally surpasses baseline models.

INDEX TERMS Pre-trained models, quantum language models, residual connection.

I. INTRODUCTION

Quantum-inspired language models have garnered attention due to their capacity for higher-order semantic interaction and model interpretability. Inspired by principles from quantum mechanics, recent research has propelled artificial intelligence into a new frontier. In quantum information theory, the fundamental concept of quantum computing involves leveraging exponential Hilbert space, mirroring the core approach of machine learning operating within configuration space [1], [2]. Building upon this foundational paradigm, quantum-inspired algorithms find application in Computer Vision (CV) [3] and Machine Learning (ML) [4], [5], [6]. These algorithms adopt mathematical formulations rooted in quantum theory. Additionally, it is noteworthy that researchers have systematically developed quantum-inspired language modeling algorithms in the domain of natural language processing.

In the field of machine learning, neural networks are generally regarded as “black boxes” because their internal operations are usually not directly observable and understandable. Quantum mechanics is the most accurate physical law that describes the world so far. Researchers hope to build transparent and ex post facto explainable networks from quantum physics to simulate human language. Some recent studies have shown that there are quantum-like phenomena in human cognition [7], especially in terms of language comprehension [8]. Intuitively, a sentence can be regarded as

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

a physical system with multiple words (like particles), which are usually multi-semantic (superimposed) and strongly correlated (entangled).

Van Rijsbergen pioneered the integration of information retrieval models into the mathematical framework of quantum mechanics within Hilbert space [9]. Subsequently, Sordoni et al. introduced the Quantum Language Model (QLM), a quantum probability-based model for implementation in Information Retrieval (IR) [10]. Building on QLM, Basile and Tamburini presented a quantum language model tailored for speech recognition [11]. Seeking broader applicability, Zhang et al. proposed a Neural Network based Quantum-like Language Model (NNQLM) designed for Question Answering (QA) [12]. Drawing inspiration from the quantum interference effect in retrieval processes, Jiang et al. employed a reduced density matrix representation to guide the construction of additional evidence resulting from the interaction between matching units [13]. Utilizing complex-valued representations based on quantum probability and human language units, Li et al. and Wang et al. developed distinct mathematical frameworks reflecting higher-level semantic aspects and conjugate position information, respectively [14], [15]. These frameworks employ a shared density matrix representation to calculate the quantum expected value of observable joint question and answer pairs, yielding a similarity matching score [16]. In the realm of recommendation systems, Niu and Hou proposed a novel textual representation incorporating global second-order feature interaction information to facilitate learning feature interactions in advertising recommendation [17].

Quantum-inspired language models model finer-grained semantic interactions in higher-order Hilbert spaces. However, previous methods usually capture semantic features based on context-free word vectors such as Word2Vec and GloVe. Therefore, how to effectively integrate the quantum-inspired language model and the pre-trained model and make them function under the fine-tuning paradigm of the pre-trained model has become a key issue for the further development of the quantum-inspired language model.

In the field of NLP, model training is often a slow process. Quantum-inspired language models use density matrices instead of word vectors, which is more computationally intensive. Therefore, BERT-based pre-trained models are considered. However, unlike the classical training process, quantum-inspired neural networks have better effects on modeling languages. However, in pre-trained models, density matrix representation often disrupts some of the information contained in BERT. Therefore, inspired by ODE, we use residuals to supplement the missing information to achieve better results.

In this paper, we introduce the BERT-Residual quantum language model, which is inspired by the multi-step methods of ordinary differential equations (ODE), as shown in Fig.1. This model employs the density matrix to capture high-order semantic interaction features that are not addressed in the BERT modeling process, thereby generating sentence representations and performing the first step of Residuals. Subsequently, quantum measurement is applied to these sentence representations, followed by a second residual connection with the BERT layer. This multi-step residual connection method effectively integrates the strengths of BERT representations and quantum density matrix representations, enhancing representation learning. Experimental results demonstrate that our proposed method consistently outperforms baseline models in text classification benchmarks.

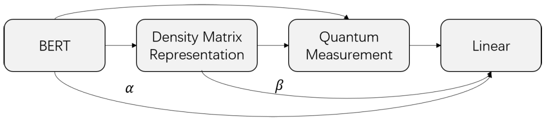

FIGURE 1. Schematic diagram of the model framework. Inspired by the multi-step method of ordinary differential equations, it fully combines the semantic information between pre-trained language models and quantum-inspired models.

- We combine BERT with quantum-inspired language models in an easy-to-compute way.
- We introduced the multi-step method of ordinary differential equations into the quantum-inspired language models based on BERT representation, achieving an effective combination of coarse-grained semantics and fine-grained semantic information.
- Experiments prove that our BRQLM model surpasses baseline models on multiple text classification datasets.

II. RELATED WORKS

A. WORD EMBEDDING-BASED CNN AND QUANTUM-INSPIRED MODEL

Text classification tasks can be implemented based on various improved versions of standard neural networks [18]. At the same time, some typical improved models based on CNN include: DPCNN [19] proposed by Johnson and Zhang, VDCNN [20] proposed by Conneau et al. and CharTextCNN [21] proposed by Zhang et al. Additionally, quantum-inspired classification models like QPDN [22] proposed by Wang et al., QICNN [23] proposed by Shi et al., and QRNN [24] proposed by Li et al. have emerged, providing innovative approaches to tackling classification tasks. Pretrained word embeddings are widely used in various works as an important component of downstream models, and they can significantly improve performance relative to embeddings learned from scratch. Although many state-of-the-art results have been achieved, further improvements in model performance face limitations such as ambiguity and task-specific structural dependencies. Although pre-trained word embeddings can capture the semantic information and contextual relationships of words, the model may be confused when faced with ambiguity. The same word can have different meanings in different contexts, and pre-trained word embeddings struggle to accurately capture this ambiguity.

This semantic ambiguity can lead to model errors when handling context-dependent tasks.

B. BERT

BERT is a language representation model that represents **B**idirectional **E**ncoder **R**epresentation from **T**ransformer [25]. BERT aims to pre-train deep bidirectional representations from unlabeled text by jointly conditioning left and right context in all layers. As a result, pre-trained BERT models can be fine-tuned with an additional output layer to create state-of-the-art models for a wide range of tasks such as question answering and language inference without requiring extensive modifications to the task-specific architecture.

Bert is widely used in the field of natural language processing. In the field of sentiment analysis, Gao et al. proposed TD-BERT [26], which implements three goal-related variants of the BERT base model, locating the output in the target item and an optional sentence with a built-in goal. And the TD-BERT model achieves new state-of-the-art performance compared to traditional approaches. Yu et al. proposed a text classification model BERT4TC [27] based on BERT, which converts the classification task into a binary sentence pair task by constructing auxiliary sentences, aiming to solve the problem of limited training data and task awareness issues. In addition, BERT has demonstrated its representative performance in document classification [28], dynamic fusion hierarchical graph methods combined with contextual node embedding [29], and automated preparation of the classification of various book components [30]. Recently, BERT-based related models have also shown improvements in the performance of small sample learning (FSL) [31].

C. RESIDUAL CONNECTION

The residual connection technique addresses the issue of gradient vanishing in deep networks [32]. When gradients saturate, it ensures network stability through identity mapping principles. Drawing from the theory of differential dynamics, residual connections can be viewed as a discretized application of the first-order Euler solver [33].

Models such as residual networks, recurrent neural network decoders, and normalizing flows construct complex transformations by sequentially applying transformations to a hidden state. This process can be expressed as:

$$h_{t+1} = h_t + f(h_t, \theta_t) \tag{1}$$

where $t \in \{0, \ldots, T\}$ and $h_t \in \mathbb{R}^D$. These iterative updates can be interpreted as an Euler discretization of a continuous transformation, as suggested by Haber and Ruthotto [34] and Haber et al. [35].

Higher-order solvers, such as the second-order Runge-Kutta method, mitigate approximation errors stemming from identity mapping [36], [37], and exhibit enhanced capability in handling continuously sampled data [38].

III. QUANTUM THEORY PRELIMINARIES

Quantum mechanics is now the most accurate physical theory that describes the world. It fundamentally changes human understanding of the material structure and interactions, and is widely used in various fields of science. Quantum mechanics is defined by a complete set of mathematical formalism. This section will briefly recapitulate some basic notations and concepts of quantum theory used in this work.

A. QUANTUM PROBABILITY

In quantum theory, physical observable are represented by compact Hermition operators in Hilbert space $\mathcal{H}$. Given a Hermition operator A, its corresponding eigenvalues and eigenvectors satisfy the following relationship.

$$A|\psi_i\rangle = a_i|\psi_i\rangle \tag{2}$$

where a_i is the eigenvalue and ψ is the eigenvector. Density matrices can be represented by Hermitian operators. The more general quantum state space includes mixed states, that is, convex combinations of pure states. Density operator representation can uniformly represent pure states and mixed states. If a quantum state ρ consists of a convex combination of pure states $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle, \cdots, |\psi_i\rangle$, the combination coefficients are $p_1, p_2, p_3, \cdots, p_i$ respectively, then the density operator representation of ρ is defined as:

$$\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i| \tag{3}$$

where p_i represents the probability of occurrence of each pure state, and need to satisfy the normalization conditions $0 \le p_i \le 1$, $\sum_i p_i = 1$.

B. THE QUANTUM LANGUAGE MODEL

The pioneering work on quantum language models comes from [9] and [10]. This section provides a brief overview of the basic framework.

In the quantum language model, the dependencies between words are represented by quantum basic events. For each word w_i, it corresponds to an observable O_i (also called a projection operator), where $O_i = e_i e_i^\dagger$ and e_i represents the word based on one-hot vector. The Neural Network based Quantum-like Language Model (NNQLM) [12] extends this approach that embeds a word as a unit vector and a sentence as a real-valued density matrix. The distance between a pair of density matrices is achieved by extracting features of matrix multiplication. In quantum mechanics, this method is called measurement. The formula for calculating the probability of words and sentences is given by Gleason's theorem [39].

$$p(m) = tr(O_i\rho) \tag{4}$$

where p_m represents the measurement probability of the density matrix ρ that represents each word/sentence. Multiply the probabilities obtained by measuring the quantum states corresponding to all words in the document to obtain the maximum likelihood function:

$$L_{p(m)} = \prod_{i=1}^{M} tr(O_i \rho) \tag{5}$$

Then the maximum likelihood can be expressed as:

$$\underset{\rho}{\text{maximize}} \ \log L_{p(m)} = \sum_{i=1}^{M} \log tr(O_i \rho) \tag{6}$$

In the original QLM [10], the above maximization problem is solved iteratively. This process is non-differentiable, so it cannot be performed in an end-to-end neural network architecture alone, but must rely on external processing of the neural network architecture. This also prompted us to improve this through neural network structures.

IV. BERT-RESIDUAL QUANTUM LANGUAGE MODEL

Drawing upon the theoretical foundations of quantum and semantic Hilbert spaces described in Section III, this paper introduces a novel BERT-Residual Quantum Language Model inspired by the ODE multi-step method (BRQLM). Fig. 2 illustrates the structural components of the proposed model. In this chapter, we provide a comprehensive discussion on five key aspects: basic concepts of BERT, density matrix-based sentence modeling, quantum measurement-based feature extraction, residual connection mechanism, and vector mapping layer.

A. UTILIZING BERT'S OUTPUT AS WORD EMBEDDING VECTORS

We treat the token output by BERT [25] as the word embedding, with the shape of $E \in R^{n \times |V| \times d}$, where n is the length of the input data, $|V|$ is the size of the word list, which is the number of words, and d is the dimension in which words are embedded in BERT. By doing so, the mapping from the original sentence to the word vector can be obtained. Due to the existence of the self-attention mechanism in BERT, each word is updated by all other words in this sentence. Therefore, we believe that the word vector output by BERT is more in line with the concept of superposition in Sec. III. In this chapter, the words ω output by BERT are post-processed as follows, and their length and direction are associated with different physical meanings: the norm of the vector represents the relative weight of the words, while the direction of the vector is considered as the superimposed words. Its formalization is:

$$|\omega\rangle = \frac{\vec{\omega}}{\|\vec{\omega}\|}, \pi(\omega) = \|\vec{\omega}\|, \tag{7}$$

where $\|\vec{\omega}\|$ represents the L2 norm of $\vec{\omega}.\pi(\omega)$ is used to calculate the weight of words in a sentence, which will be explained in the following text. It is worth noting that the normalized weights of specific words in the method proposed in this article are not static, but are adaptively updated during the training phase.

B. DENSITY MATRIX-BASED SENTENCE MODELING

The previous chapter has already discussed modeling sentences as a combination of word vectors output by BERT. Next, first calculate the projection, which is the word vector ω_i embedded by the word i. The subspace spanning is calculated from the outer product as Eq. 8, as shown at the bottom of the next page.

Next, add weights to each word in the sentence. In the Eq. 3, each word is given the same weight, which empirically does not hold because each word has a different weight in the sentence. In this article, the L2 norm of the word vector is used as the relative weight of the word in this sentence, which can be updated during training. To some extent, L2 norm is a measure of the semantic richness of a word, meaning that the longer the vector, the richer its meaning, and the greater its weight in the sentence, as shown in Fig. 3.

The density matrix at the sentence level is calculated as follows Eq. 9, as shown at the bottom of the next page.

C. QUANTUM MEASUREMENT-BASED FEATURE EXTRACTION

In the field of quantum information, some people attempt to estimate quantum states through a series of measurement results. Inspired by these works, this article introduces a measurement method based on trainable word vectors to extract density matrix features.

Suppose a sentence is represented as a density matrix ρ. Apply a set of measurement operators $\{|v_k\rangle\}_{k=1}^{K}$ to this density matrix, measured using Eq. 4, can generate k probability values, which are combined into a vector p, where $p_k = \langle v_k | \rho | v_k \rangle, k \in \{1, \ldots, K\}$. Through this method, a set of measurement operators can be obtained to observe the results of a sentence, and this result can be passed as a feature into the subsequent components of the model.

In the selection of measurement operators, this article uses word vectors as measurement operators. Based on experience, summarize some high-frequency complaint reasons and segment the reasons into words. Use the word embedding method mentioned above to embed words. Due to the different number of word segmentation due to complaints, this article adds and normalizes the word embedding into a unit vector using L2 normalization. And transformed into trainable measurement operators through outer product. This is empirically feasible because the probability obtained by using word vectors as measurement operators can be seen as calculating the weight of different reasons in the dialogue text, which can be passed back to other components as features.

D. RESIDUAL CONNECTION MECHANISM

Residual connection is inspired by the first-order Euler numerical solution of ordinary differential equation theory, which is usually used to avoid the gradient vanishing problem in deep neural network training, and can also be used for the problem of shallow information forgetting.

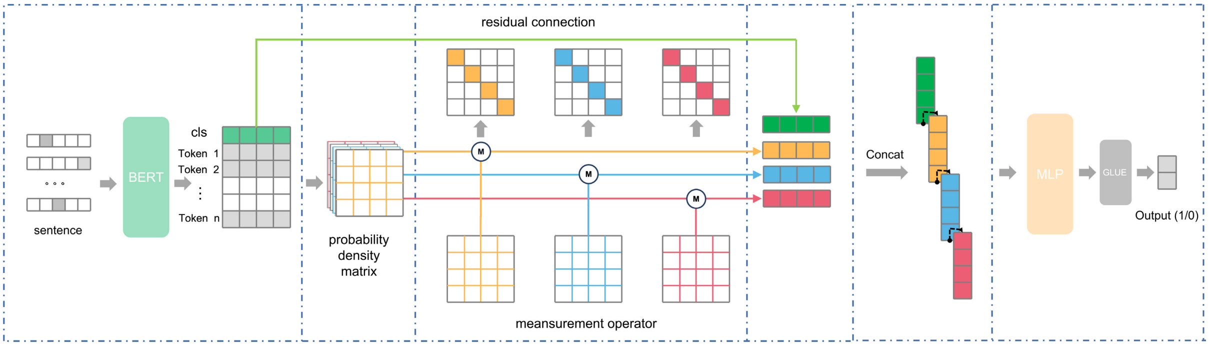

FIGURE 2. The model architecture diagram of the BERT-residual quantum language model employs BERT's word vectors, establishes a residual connection with the cls vector, and constructs a probability density matrix incorporating the remaining tokens. Trainable measurement operators are utilized for measurement, and the diagonal elements are connected to the cls vector and passed through a multi-layer perceptron (MLP) for the final classification.

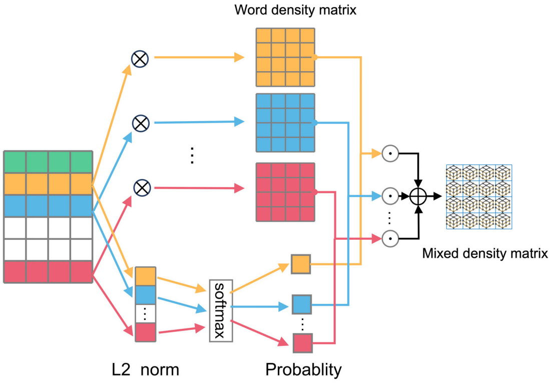

FIGURE 3. The architecture of sentence level hybrid components for generatingsentence level density matrix. ⊙ represents a matrix that multiplies numbers by each element and ⊗ represents the outer product of a vector.

Therefore, in our framework, in order to further strengthen the shallow BERT features and density matrix features, this paper is inspired by the multi-step numerical connection algorithm of ordinary differential equations, and adds the shallow output to the input of the next few layers respectively. While optimizing the gradient, it effectively integrates the high-order feature interaction advantages of the quantum density matrix with the capabilities of the pre-trained model. The specific formula is as follows:

$$\begin{aligned} output &= \mathrm{MLP}(\alpha \times \mathbf{B} + \beta \times \mathbf{D} + \mathbf{M}) \\ \mathbf{M} &= \mathrm{Density}(\mathbf{B}) \\ \mathbf{D} &= \mathrm{Measure}(\mathbf{B} + \mathbf{M}) \end{aligned} \tag{10}$$

Among them, α and β are trainable weight scalars, which effectively regulate the influence of the [CLS] vector of BERT **B** and **D** on the MLP classifier. Density() refers to the density matrix module, and Measure() refers to the quantum measurement module.

E. VECTOR MAPPING LAYER

Transfer the obtained feature vectors to the vector mapping layer for multi category classification.

V. EXPERIMENTAL DATA AND SETUP

A. DATASET

The dataset used in this study is obtained from the USPTO patents [40], consisting of 4,000 patents. Each data entry includes fields such as application number, actual content, abstract, IPC classification, and application year. The "actual" field is manually curated by reading the patent abstracts and identifying their relevance to artificial intelligence (AI) technology. The main task of this paper is

$$|\omega_i\rangle\langle\omega_i| = \begin{pmatrix}\omega'_{i1}\\ \omega'_{i2}\\ \cdots\\ \omega'_{id}\end{pmatrix} \times \left(\omega'_{i1},\omega'_{i2},\cdots,\omega'_{id}\right) = \begin{bmatrix} \left(\omega'_{i1}\right)^2 & \omega'_{i1}\omega'_{i2} & \cdots & \omega'_{i1}\omega'_{id}\\ \omega'_{i2}\omega'_{i1} & \left(\omega'_{i2}\right)^2 & \cdots & \omega'_{i2}\omega'_{id}\\ \vdots & & & \\ \omega'_{id}\omega'_{i1} & \omega'_{id}\omega'_{i2} & \cdots & \left(\omega'_{id}\right)^2\end{bmatrix} \tag{8}$$

$$\rho = \sum_i^l p(\omega_i)\,|\omega_i\rangle\langle\omega_i| = \begin{bmatrix} \sum_i p(\omega_i)\left(\omega'_{i1}\right)^2 & \sum_i p(\omega_i)\omega'_{i1}\omega'_{i2} & \cdots & \sum_i p(\omega_i)\omega'_{i1}\omega'_{id}\\ \sum_i p(\omega_i)\omega'_{i2}\omega'_{i1} & \sum_i p(\omega_i)\left(\omega'_{i2}\right)^2 & \cdots & \sum_i p(\omega_i)\omega'_{i2}\omega'_{id}\\ \vdots & & \cdots & \\ \sum_i p(\omega_i)\omega'_{id}\omega'_{i1} & \sum_i p(\omega_i)\omega'_{id}\omega'_{i2} & \cdots & \sum_i p(\omega_i)\left(\omega'_{id}\right)^2\end{bmatrix} \tag{9}$$

to determine the relevance of a patent to AI technology using the natural language in the abstract section.

B. EXPERIMENTAL SETTING

For all tasks, we implement our model with Pytorch-1.10, and train them on two NVIDIA Tesla P40 GPU. All the weight parameters are initialized with Xavier [41]. As for the learning method, we use the AdamW optimizer [42]. The learning rate is 0.006 and the batch size is 32. We perform dropout after each layer, except input and output ones, and the rate usually is set to 0.1 or 0.2. The number of measurement operators is set as $\{8,16,32,64\}$.

To conduct model training, validation, and testing, we partitioned the 4,000 data entries into training, validation, and test sets following an 8:1:1 ratio. The statistical distribution of each label in the three sets is presented in Table 1.

TABLE 1. Distribution of labels in training, validation, and test sets.

dataset	Label 0	Label 1
Training set	2558	642
Validation set	320	80
Test set	320	80
Total	3198	802

During the analysis of the dataset, it was observed that the dataset exhibits class imbalance, with a ratio of 8:2 between label 0 and label 1. Such an imbalanced dataset can pose challenges for model training. In order to comprehensively assess the capabilities of the proposed model, a resampling technique was employed to address this issue. Specifically, a 2x oversampling technique was applied to label 1, resulting in the creation of a new dataset. In this process, the data instances belonging to label 1 were duplicated twice, leading to a new dataset with a balanced ratio of 1:1 between label 0 and label 1. By utilizing these two datasets, the objective is to thoroughly evaluate the performance of the proposed model under both balanced and imbalanced sample conditions. These steps were undertaken to ensure a comprehensive assessment of the model's capabilities in handling both balanced and imbalanced datasets.

C. BASELINES

We use BERT and fully connected networks [25] as the main baselines, and compare their performance with deep neural networks based on BERT+CNN (CNN_{max} and CNN_{ave}) [36], BERT+LSTM (LSTM) [43], BERT+Attention (Attention) mechanisms [44], and other architectures. To verify the effectiveness of our proposed quantum inspired module, we froze the Bert module to exclude the impact of the Bert model.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

A. MAIN EXPERIMENT

Table 2 shows the experimental results on the dataset presented in this work. We compared the proposed BRQLM with other baseline models under two settings: original data and resampled data.

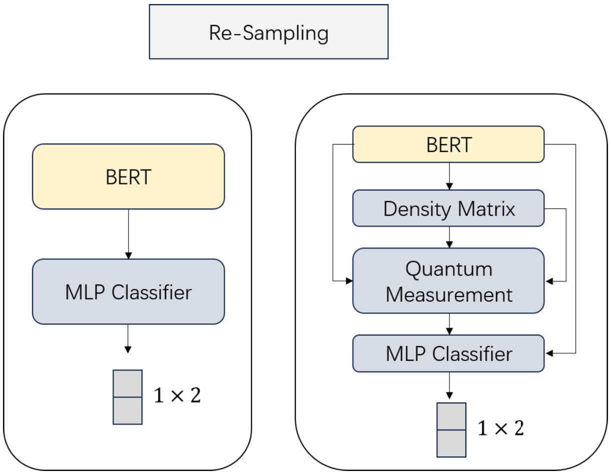

FIGURE 4. The picture on the left refers to the model framework of the baseline, and the picture on the right refers to the framework of the proposed model. Among them, re-sampling refers to 1:1 sampling of data according to categories.

In the original data distribution, compared to the baseline of BERT + MLP, the proposed model has significantly improved the ACC, Recall, F1 and AUC of the data set. Among them, the AUC indicator increased by 22.2%. This fully demonstrates that the proposed model can more fully mine the fine-grained information in data representation, thereby greatly improving the accuracy of model predictions. This indicates that the density matrix representation based on the multi-step method can not only fully preserve the semantic information implied in BERT but also model higher-order feature interactions. Compared to the best-performing overall CNN_{avg} model, we demonstrated superior performance in terms of Recall and AUC metrics, with an improvement of 2.52%.

On the resampling setting, compare with all baseline models, BRQLM achieves significant improvements in all five performance metrics. This shows that whether it is balanced data or uneven data distribution, the proposed model can capture the interactive relationships in the text that are more consistent with human cognition, thereby achieving more accurate text understanding. In terms of the Accuracy (Acc) metric, we achieved a 1.49% improvement compared to the CNN_{ave} model and a 4.61% improvement compared to the BERT+MLP baseline.

In Table 3, we present comparative experiments conducted on the GLUE benchmark, which demonstrate the effectiveness of our proposed method. Specifically, compared to BERT+CNN, BERT+LSTM, and BERT+Attention models, BRQLM achieves an improvement of over 3.33% in the accuracy (ACC) evaluation metric for the MRPC dataset. Notably, on the RTE dataset, BRQLM enhances the ACC metric by 18.17% and the F1-score metric by 35.79% compared to the CNN model. Furthermore, in comparison

TABLE 2. Experimental results chart.

Model	Acc	Precision	Recall	F1 score	AUC
CNN$_{max}$	0.9163	0.8663	0.8569	0.8615	0.8569
CNN$_{ave}$	**0.9187**	0.8719	0.8585	**0.8649**	0.8585
LSTM	0.8725	0.7928	0.7868	0.7897	0.7868
Attention	0.9163	**0.8752**	0.8417	0.8570	0.8417
BERT+MLP	0.8775	0.8455	0.7187	0.7581	0.7187
BRQLM	0.9100	0.8448	**0.8785**	0.8600	**0.8785**
CNN$_{max}$(re-sampling)	0.8973	0.8949	0.8976	0.8960	0.8976
CNN$_{ave}$(re-sampling)	0.8982	0.8959	0.8979	0.8968	0.8979
LSTM(re-sampling)	0.8839	0.8816	0.8864	0.8830	0.8864
Attention(re-sampling)	0.9045	0.9026	0.9085	0.9038	0.9085
BERT+MLP (re-sampling)	0.8714	0.8711	0.8668	0.8686	0.8668
BRQLM (re-sampling)	**0.9116**	**0.9095**	**0.9117**	**0.9105**	**0.9117**

TABLE 3. ACC and F1 results comparison for different datasets.

	Acc				**F1 score**			
Dataset	**CNN**	**LSTM**	**Attention**	**BRQLM**	**CNN**	**LSTM**	**Attention**	**BRQLM**
MRPC	0.6886	0.6892	0.6863	**0.7122**	0.8071	0.8064	0.8064	**0.8161**
QNLI	0.7869	0.7713	0.7639	**0.7921**	0.8022	0.7821	0.7817	**0.8192**
QQP	0.8107	**0.8263**	0.8025	0.7626	0.7462	**0.7727**	0.7534	0.686
RTE	0.5156	0.5703	0.5234	**0.6093**	0.4876	0.5378	0.5793	**0.6621**
SST-2	0.8587	0.875	0.8796	**0.8916**	0.8644	0.88	0.8828	**0.9012**
WNLI	0.5625	0.4687	0.484	**0.5625**	0	0.3461	0.1951	**0.3**

TABLE 4. Results of ablation experiment.

Model	Acc	Precision	Recall	F1 score	AUC
w / o residual	0.8113	0.4056	0.5000	0.4479	0.5000
w / o quantum modules	0.8775	0.8455	0.7187	0.7581	0.7187
BRQLM	0.9100	0.8448	0.8785	0.8600	0.8785
w / o residual(re-sampling)	0.8402	0.8380	0.8368	0.8374	0.8368
w / o quantum modules(re-sampling)	0.8714	0.8711	0.8668	0.8686	0.8668
BRQLM (resampling)	**0.9116**	**0.9095**	**0.9117**	**0.9105**	**0.9117**

to the Attention model, BRQLM increases the ACC metric by 16.22% and the F1-score metric by 53.77% on the WNLI dataset.

In Table 3, we present the results of our comparative experiments conducted on various datasets, showcasing the effectiveness of our proposed BRQLM model in contrast to traditional models such as CNN, LSTM, and Attention.

Considering computational parallelism, the quantum density matrix and quantum measurement module are both based on matrix operations, so they can be parallelized using GPUs. Specifically, the computational complexity of the density matrix is $O(d^2 \times N)$, and the computational complexity of quantum measurement is $O(d^2 \times M)$. In this context, d represents the size of the hidden layer, N denotes the size of the sequence length, and M indicates the number of measurement operators. Our experiments ran smoothly on a P40 machine with 23GB of memory.

B. ABLATION TEST

The efficacy of different constituents in quantum models has been extensively examined in prior research, as evidenced by notable studies conducted by [13] and [14], among others. In this context, we focus on verifying the performance improvement achieved by introducing a residual module. To this end, we implemented a quantum-inspired BERT network without residual connections and conducted experiments. The experimental results are presented in Table 4.

Our experiments demonstrate that, while BERT's text embeddings contain sufficient information, the quantum language model's ability to capture finer-grained features is significantly enhanced when paired with a residual mechanism. This mechanism is crucial for mitigating network issues associated with quantum computing, particularly the risk of overfitting.

Additionally, we also examined the impact of ablating the quantum module itself. Removing the quantum module resulted in a noticeable decline in the model's performance, indicating its vital role in enhancing the model's expressive power and feature extraction capabilities. This further emphasizes that both the residual mechanism and the quantum module are essential components that contribute positively to the overall performance of the model.

The term "w/o residual" refers to a configuration that eliminates residual connections. In our experiments with the original dataset, we observed a notable decline in model performance, primarily due to the challenges of class imbalance and overfitting, which resulted in a 41.21% decrease in the F1 score. Conversely, when employing a resampling setup, the inclusion of residual connections led to a 7.31% improvement in the model's F1 score. Furthermore, w / o quantum modules resulted in decreases in F1 scores of 10.19% and 4.19%, respectively. These findings indicate that both modules make a significant positive contribution to model performance.

VII. CONCLUSION

This article aims to use the BERT model to capture semantic relationships through contextual information, so that word vectors can better capture the semantic information of the text. Next, we will encode the word vectors output by BERT into a sentence density matrix. Extract the features of the sentence density matrix through trainable measurement operators, retain the original sentence features through residual networks, and then use a simple concatenation and fusion mechanism, and achieve the binary classification task of sentences through fully connected neural networks. In the task of patent classification, we conducted a series of comparative and ablation experiments on our model without resampling or resampling twice the dataset, ultimately verifying the effectiveness of our BERT residual quantum language model.

Specifically, the paper enhances the fine-grained semantic understanding capability of the BERT model by incorporating quantum density matrix and quantum measurement. It also employs a discretized form of second-order numerical solutions to ordinary differential equations for network modeling, thereby avoiding gradient issues caused by quantum heuristic calculations and improving the expressive power of deep language models. However, the complexity of the density matrix limits the application of this method in larger-scale pre-trained language models. Therefore, we will further explore the use of reduced density matrices, which have lower computational complexity, to enhance the capability of pre-trained language models. At the same time, we will consider the use of Physics-Informed Neural Networks [45], [46], in such a framework and the particularities to be used for quantum language model [47]. We will also consider the efficiency and practical applications of quantum theory in large machine learning models [48].

REFERENCES

[1] M. Schuld and N. Killoran, "Quantum machine learning in feature Hilbert spaces," *Phys. Rev. Lett.*, vol. 122, no. 4, Feb. 2019, Art. no. 040504.

[2] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge, MA, USA: Cambridge Univ. Press, 2010.

[3] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, "Quantum entanglement in deep learning architectures," *Phys. Rev. Lett.*, vol. 122, no. 6, Feb. 2019, Art. no. 065301.

[4] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," *Nature*, vol. 549, no. 7671, pp. 195–202, Sep. 2017.

[5] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning," *Nature Commun.*, vol. 12, no. 1, pp. 1–9, May 2021.

[6] X. Gao, Z.-Y. Zhang, and L.-M. Duan, "A quantum machine learning algorithm based on generative models," *Sci. Adv.*, vol. 4, no. 12, Dec. 2018, Art. no. eaat9004.

[7] D. Aerts and S. Sozzo, "Quantum entanglement in concept combinations," *Int. J. Theor. Phys.*, vol. 53, no. 10, pp. 3587–3603, Oct. 2014.

[8] P. Bruza, K. Kitto, R. D. McEvoy, and C. L. McEvoy, "Entangling words and meaning," in *Proc. 2nd Quantum Interact. Symp. (QI)*, Jan. 2008, pp. 118–124.

[9] C. J. Van Rijsbergen, *The Geometry of Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[10] A. Sordoni, J.-Y. Nie, and Y. Bengio, "Modeling term dependencies with quantum language models for IR," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 653–662.

[11] I. Basile and F. Tamburini, "Towards quantum language models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1840–1849.

[12] P. Zhang, J. Niu, Z. Su, B. Wang, L. Ma, and D. Song, "End-to-end quantum-like language models with application to question answering," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–8.

[13] Y. Jiang, P. Zhang, H. Gao, and D. Song, "A quantum interference inspired neural matching model for ad-hoc retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 19–28.

[14] Q. Li, B. Wang, and M. Melucci, "CNM: An interpretable complex-valued network for matching," 2019, *arXiv:1904.05298*.

[15] B. Wang, D. Zhao, C. Lioma, Q. Li, P. Zhang, and J. Grue Simonsen, "Encoding word order in complex embeddings," 2019, *arXiv:1912.12333*.

[16] Q. Zhao, C. Hou, C. Liu, P. Zhang, and R. Xu, "A quantum expectation value based language model with application to question answering," *Entropy*, vol. 22, no. 5, p. 533, May 2020.

[17] T. Niu and Y. Hou, "Density matrix based convolutional neural network for click-through rate prediction," in *Proc. 3rd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2020, pp. 46–50.

[18] Y. Chen, "Convolutional neural network for sentence classification," M.S. thesis, Dept. Syst. Des. Eng., Univ. Waterloo Canada, 2015.

[19] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 562–570.

[20] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," 2016, *arXiv:1606.01781*.

[21] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.

[22] B. Wang, Q. Li, M. Melucci, and D. Song, "Semantic Hilbert space for text representation learning," in *Proc. World Wide Web Conf.*, May 2019, pp. 3293–3299.

[23] S. Shi, Z. Wang, G. Cui, S. Wang, R. Shang, W. Li, Z. Wei, and Y. Gu, "Quantum-inspired complex convolutional neural networks," *Appl. Intell.*, vol. 52, no. 15, pp. 17912–17921, Dec. 2022.

[24] Y. Li, Z. Wang, R. Han, S. Shi, J. Li, R. Shang, H. Zheng, G. Zhong, and Y. Gu, "Quantum recurrent neural networks for sequential learning," *Neural Netw.*, vol. 166, pp. 148–161, Sep. 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[27] S. Yu, J. Su, and D. Luo, "Improving BERT-based text classification with auxiliary sentence and domain knowledge," *IEEE Access*, vol. 7, pp. 176600–176612, 2019.

[28] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for document classification," 2019, *arXiv:1904.08398*.

[29] A. Onan, "Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 7, Jul. 2023, Art. no. 101610.

[30] P. Wulff, L. Mientus, A. Nowak, and A. Borowski, "Utilizing a pretrained language model (BERT) to classify preservice physics teachers' written reflections," *Int. J. Artif. Intell. Educ.*, vol. 33, no. 3, pp. 439–466, Sep. 2023.

[31] W. Liao, Z. Liu, H. Dai, Z. Wu, Y. Zhang, X. Huang, Y. Chen, X. Jiang, D. Liu, D. Zhu, S. Li, W. Liu, T. Liu, Q. Li, H. Cai, and X. Li, "Mask-guided BERT for few-shot text classification," *Neurocomputing*, vol. 610, Dec. 2024, Art. no. 128576. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122401347X

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31. Curran Associates, 2018.

[34] E. Haber and L. Ruthotto, "Stable architectures for deep neural networks," *Inverse Problems*, vol. 34, no. 1, Dec. 2017, Art. no. 014004, doi: 10.1088/1361-6420/aa9a90.

[35] E. Haber, L. Ruthotto, E. Holtham, and S.-H. Jun, "Learning across scales—Multiscale methods for convolution neural networks," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 1–7. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/0

[36] M. Zhu, B. Chang, and C. Fu. (2019). *Convolutional Neural Networks Combined With Runge–Kutta Methods*. [Online]. Available: https://openreview.net/forum?id=HJNJws0cF7

[37] B. Li, Q. Du, T. Zhou, S. Zhou, X. Zeng, T. Xiao, and J. Zhu, "ODE transformer: An ordinary differential equation-inspired model for neural machine translation," 2021, *arXiv:2104.02308*.

[38] T. Demeester, "System identification with time-aware neural sequence models," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 4, pp. 3757–3764.

[39] A. M. Gleason, "Measures on the closed subspaces of a Hilbert space," in *The Logico-Algebraic Approach To Quantum Mechanics: Volume I: Historical Evolution*. Berlin, Germany: Springer, 1975, pp. 123–133.

[40] M. Miric, N. Jia, and K. G. Huang, "Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents," *Strategic Manage. J.*, vol. 44, no. 2, pp. 491–519, Feb. 2023.

[41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Mach. Learn. Res.*, vol. 9, Y. W. Teh and M. Titterington, Eds., May 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[43] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM neural network for text classification," 2015, *arXiv:1511.08630*.

[44] X. Sun and W. Lu, "Understanding attention for text classification," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3418–3428.

[45] J. de Curtò and I. de Zarzà, "Hybrid state estimation: Integrating physics-informed neural networks with adaptive UKF for dynamic systems," *Electronics*, vol. 13, no. 11, p. 2208, Jun. 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/11/2208

[46] F. S. Costabal, S. Pezzuto, and P. Perdikaris, "Δ-PINNs: Physics-informed neural networks on complex geometries," *Engineering Appl. Artif. Intell.*, vol. 127, Jan. 2022, Art. no. 107324.

[47] C. Trahan, M. Loveland, and S. Dent, "Quantum physics-informed neural networks," *Entropy*, vol. 26, no. 8, p. 649, Jul. 2024. [Online]. Available: https://www.mdpi.com/1099-4300/26/8/649

[48] J. Liu, M. Liu, J.-P. Liu, Z. Ye, Y. Wang, Y. Alexeev, J. Eisert, and L. Jiang, "Towards provably efficient quantum algorithms for large-scale machine-learning models," *Nature Commun.*, vol. 15, no. 1, p. 434, Jan. 2024.

SHAOHUI LIANG received the B.E. degree in computer science and technology from Shijiazhuang University of Economics, in 2006, and the M.E. degree in computer technology from Tianjin University, in 2014. He is currently working as an Associate Professor with Tianjin Renai College. His research interests include quantum artificial intelligence and machine learning.

YINGKUI WANG received the B.E. and M.E. degrees in statistics from Sichuan University, Sichuan, China, in 2005 and 2008, respectively, and the Ph.D. degree from Tianjin University, China, in 2021. He is currently working as a Lecturer with the School of Computer Science and Technology, Tianjin Renai College. His current research interests include community detection and deep learning.

SHUXIN CHEN received the B.E. degree in computer science and technology from Qiqihar University, in 2001, the M.E. degree in thermal engineering from Harbin University of Science and Technology, in 2010, and the Ph.D. degree in optical engineering from Harbin Institute of Technology, in 2019. She is currently working as a Professor with the School of Computer Science and Technology, Tianjin Renai College. Her current research interests include computer simulation, big data analysis, and quantum computing.