Comparative Study of Traditional Machine Learning and Quantum Computing in Natural Language Processing: A Case Study on Sentiment Analysis

Li Yang
Faculty of Data Science
Musashino University
Tokyo, Japan
s2122055@stu.musashino-u.ac.jp

Thatsanee Charoenporn
Asia AI Institute,
Faculty of Data Science,
Musashino University
Tokyo, Japan
thatsane@musashino-u.ac.jp

Virach Sornlertlamvanich
Asia AI Institute,
Faculty of Data Science
Musashino University
Tokyo, Japan
virach@musashino-u.ac.jp

***Abstract*—This research paper primarily investigates the application and performance of traditional machine learning and quantum computing in Natural Language Processing (NLP), with a focus on sentiment analysis tasks. By comparing the accuracy, efficiency, and scalability of these two technologies, the study aims to reveal the potential of quantum computing in handling complex NLP tasks and to provide data support for future technology choices and research directions. The paper also details the use of IMDB and NLTK movie review datasets for experiments and discusses the experimental design, performance evaluation results, and technical challenges faced by both methods.**

Keywords—NLP, Quantum Computing, Sentiment Analysis, Machine Learning, Performance Comparison

I. Introduction

Natural Language Processing (NLP) is a core branch of artificial intelligence that is dedicated to enabling computers to understand and generate human language. With the development of big data and increasingly sophisticated algorithms, NLP technology has been widely applied in various scenarios such as search engines, chatbots, and sentiment analysis. Particularly in sentiment analysis, by recognizing the emotional tendencies in texts, it helps businesses gain insights into consumer emotions, improve products and services, and has become an indispensable tool in business intelligence and social media monitoring.

Although traditional machine learning methods have achieved significant success in sentiment analysis, these methods generally rely on large amounts of data and complex feature engineering, and face challenges of computational efficiency and resource consumption when dealing with unstructured, large-scale textual data. Recently, quantum computing, as a potential breakthrough technology, is considered to have unique capabilities in handling complex NLP tasks due to its advantages in parallel processing and efficient algorithms. Therefore, exploring and comparing the applications of traditional machine learning and quantum computing methods in sentiment analysis tasks can help assess the practical application prospects of quantum computing in the NLP field.

II. Research Objectives and Contributions

In traditional computers, a "bit" is the basic unit of information and is typically used to represent a binary digit, 0 or 1. Each bit can be in one of two states, commonly represented by the numbers 0 and 1. This representation is based on Boolean logic, where 0 generally signifies false, and 1 signifies true. The arithmetic logic unit (ALU), such as a CPU within a computer, performs various logical and arithmetic operations, which are essentially carried out on bits. For example, the addition of two numbers at the most fundamental level is performed at the bit level. Logic gates within the computer, such as AND, OR, and NOT, use bits to perform basic logical operations. For instance, an AND gate only outputs 1 when both of its inputs are 1; otherwise, it outputs 0.

A quantum bit, or qubit, is the fundamental unit of information in quantum computing, analogous to the bit in traditional computing. However, qubits leverage properties of quantum mechanics, such as superposition and entanglement, offering new mechanisms and potential for processing information. Unlike traditional bits that must be either 0 or 1 at any given moment, a qubit can exist in both 0 and 1 states simultaneously, a state known as superposition. Specifically, the state of a qubit can be represented as $\alpha|0> + \beta|1>$, where $|0>$ and $|1>$ are the quantum states' basis states, and α and β are complex probability amplitudes that determine the probability magnitude and phase of the qubit being in the respective states. When the quantum states of two or more qubits become inseparable due to interactions, they are said to be entangled. Entanglement is a powerful resource in quantum computing, allowing quantum computers to surpass the capabilities of traditional computers in some computations. However, unlike traditional bits, the state of a qubit cannot be determined before measurement. Measuring a qubit causes its state to "collapse" to one of $|0>$ or $|1>$, with the probability of each outcome determined by the corresponding probability amplitude in the superposition. Due to the property of superposition, a single qubit can represent multiple potential states simultaneously. For example, two qubits can simultaneously represent four possible states (00, 01, 10, 11). As the number of qubits increases, this capability grows exponentially, whereas traditional bits increase linearly. Superposition enables quantum computers to process multiple states in a single operation, providing opportunities for certain types of parallel processing and algorithm optimization, such as the famous Shor's algorithm and Grover's algorithm. The parallelism of quantum computing theoretically allows for the simultaneous processing of a vast number of data states, particularly useful in NLP where corpora are typically very large. Quantum computers can effectively enhance processing speed and efficiency when handling semantic relationships, building language models, or performing complex semantic analyses. In quantum computing, the concept of quantum gates is essential. Quantum gates are the basic operational units in quantum computing, analogous to logic gates in traditional computing. However, quantum gates operate on qubits and can perform more complex operations, such as transforming superposition and entangled states. Quantum gates utilize principles of quantum mechanics, such as superposition and quantum entanglement, to perform basic arithmetic and logical operations. Operations of quantum gates are based on linear algebra, and each quantum gate can be represented as a unitary matrix. A unitary matrix is a special type of square matrix that satisfies the condition that the matrix multiplied by its conjugate transpose equals the identity matrix. This property ensures the reversibility of quantum computing, meaning that each quantum operation can be completely reversed. Quantum gates can be chained and combined to construct complex quantum circuits, which are used to execute various algorithms such as the quantum Fourier transform, search algorithms, and cryptographic algorithms. By combining different quantum gates, any quantum algorithm can be implemented on a quantum computer, thereby leveraging the properties of quantum mechanics to solve problems that are difficult to address with traditional computers.

This study aims to quantify the differences in accuracy, efficiency, and scalability between traditional machine learning methods and quantum computing approaches in sentiment analysis through comparative analysis. By making this comparison, we hope to reveal the potential of quantum computing in natural language processing and provide data support for future technology choices and research directions.

III. Related Research

A. *Traditional Machine Learning in Sentiment Analysis*

Sentiment analysis is an important research direction in the NLP field, aiming to identify and classify emotional states in texts. Traditional machine learning methods have been widely applied in sentiment analysis. For example, algorithms such as Support Vector Machines (SVM) and Random Forests are often used to build sentiment classification models. These models predict the emotional tendencies of texts by learning the semantics and contextual relationships of words. Additionally, deep learning methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become mainstream in sentiment analysis due to their superior ability in handling sequential data and interpreting complex semantics. These methods usually require a large amount of labeled data for training and rely on complex feature engineering to enhance model performance.

B. *Exploration of Quantum Computing in NLP*

Quantum computing is a technology that utilizes the quantum mechanical properties of quantum bits (qubits) [1] for information processing. Compared to traditional computing, quantum computing exhibits theoretical superior performance in handling certain types of problems. Although the application of quantum computing in the NLP field is still in its infancy, some preliminary studies have shown its potential. For example, quantum algorithms such as Quantum Neural Networks (QNNs) and Quantum Variational Classifiers have been used for text classification and sentiment analysis tasks [2, 3]. These studies typically explore how to encode textual data into quantum states and utilize the parallelism of quantum algorithms to speed up model training and inference processes [4]. In NLP), many problems can indeed be considered as optimization problems, including but not limited to parameter tuning of language models, text clustering, feature selection, and various forms of structural prediction problems. Quantum computing offers a potential way to accelerate the process of solving these problems, especially through quantum annealing [5] and Quantum-based optimization algorithms. Quantum annealing is a technique that utilizes quantum tunneling effects to find globally optimal solutions. It is particularly suitable for solving complex optimization problems with multiple local minima, which are common in high dimensional and complex NLP model parameter spaces. In parameter optimization of language models, quantum annealing can help the model to escape from local optima more efficiently and potentially find better global solutions, thus improving the overall performance of the model. For text clustering, quantum annealing can be used to optimize the objective function of the clustering algorithm, such as minimizing the intra-class distance while maximizing the inter-class distance. In conclusion, quantum computing demonstrates some prospective application potentials in optimization problems in NLP, but quite a few technical and practical challenges need to be overcome to realize these potentials [6]. With the development of quantum technology, we may see more innovative solutions being proposed to address these challenges afterwards.

C. *Significance of Comparative Research*

Although the application of quantum computing in NLP is still at an experimental stage, its theoretical advantages prompt researchers to explore its possible application scenarios. Comparative research on the performance of traditional machine learning and quantum computing in sentiment analysis tasks not only helps to validate the effectiveness of quantum algorithms in practical NLP applications but also reveals the performance differences between the two at the current technology level. Moreover, through such comparisons, we can further understand the potential advantages and limitations of quantum computing in processing language data, providing guidance for future algorithm design and technology improvements [7].

D. *Problem and Dataset Overview*

In today's digital age, sentiment analysis has become a crucial topic in the field of NLP. It involves the automatic detection and classification of emotional undertones in text material. This study utilizes the widely recognized IMDB movie review dataset, which contains numerous movie review texts, each labeled as positive or negative to express the reviewer's overall sentiment towards the film.

For this task, we employed the Keras library to load a preprocessed dataset that includes the most common 10,000 words. This approach simplifies the model's input requirements and helps eliminate noise data, allowing the machine learning algorithms to more effectively learn and predict the emotional tones of the reviews. The dataset is divided into a training set and a test set, with the training set used for model training and the test set for evaluating model performance.

IV. Traditional Machine Learning Methods

A. *Traditional NLP Analysis*

To ensure that the model can effectively process and analyze text data, we first preprocessed the comments in the IMDB dataset by mapping the integer indices back to their corresponding words. This step is crucial because the text in the dataset initially exists in the form of word indices, which correspond to a predefined vocabulary in the dataset. By creating a reverse mapping dictionary from indices to words, we can transform each comment's index sequence back into its original word form, producing readable text.

Next, we used scikit-learn's train_test_test_split method to randomly divide the data into training and testing sets, with 25% of the data used as the test set. This step not only helps assess the model's generalization ability but also ensures the randomness of the data, preventing biases during the model training process.

Through these preprocessing steps, we ensured the consistency and quality of the model inputs, laying a solid foundation for subsequent model training and validation.

B. Traditional Machine Learning Methods

In this study, we initially employed a traditional machine learning model based on logistic regression for sentiment analysis. Logistic regression is a widely used classification technique suitable for binary classification problems. We integrated CountVectorizer and Logistic Regression (1) through a pipeline. CountVectorizer is responsible for converting text data into a bag-of-words model, simplifying the textual data into a numerical format that the model can process.

The mathematical equation for Logistic Regression is as follows:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

The training process involved in putting the transformed text data (features) and corresponding labels (sentiment) into the logistic regression classifier. In Fig.1, to obtain more stable results, we set a high number of iterations (max_iter=500) to ensure sufficient iterations before convergence. After training the model, we evaluated its accuracy using test set data to verify the model's ability to classify unseen data.

From Fig 1, we can see that in the traditional logistic regression model, the accuracy holds a significant advantage, with correct predictions making up 87% of the total. This indicates that the model performs well on the current dataset. Although there is a 13% error rate, the overall accuracy is sufficiently high for many practical applications. This also shows that even simple models like logistic regression can provide efficient and reliable predictions under suitable conditions and with appropriate datasets.

To visually present the model's prediction results, we used the Matplotlib and Seaborn libraries to plot count graphs of the predictions. These charts display the number of comments the model predicted as positive and negative, helping us visually understand how the model tends to classify comments.

C. Quantum Computing Simulation

Data Preprocessing and Feature Simplification: Extract reviews from the NLTK movie review dataset and shuffle them randomly. Simplified feature extraction includes using the length of reviews as the only feature to reduce quantum resource demands.

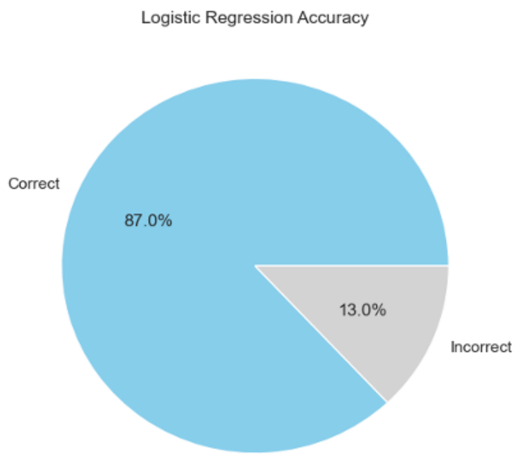

Fig. 1. Sentiment Prediction – Logistic Regression

Quantum Circuit Construction: Build a quantum circuit using Google's Cirq [8] library. The quantum circuit simulates a simple binary classifier, representing and processing information through quantum bits.

Quantum State Measurement and Classification: Predict emotional categories through the measurement results of the quantum circuit, with the final measurement state of the quantum bits determining the emotional tendency of the reviews. Fig. 2 shows the result of the RX gates and measurement gates.

In quantum computing, the Bloch sphere is a geometric representation used to visualize the state of a single qubit. A qubit's state can be a quantum superposition of the 0 state (|0>) and the 1 state (|1>) , and this superposition state can be represented by a point on the Bloch sphere.

The Bloch sphere offers an intuitive way to visualize the evolution of a qubit's state, especially when quantum gates, like rotation gates, are applied.

Any quantum state of a qubit can be expressed as:

$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle \tag{2}$$

$\cos\left(\frac{\theta}{2}\right)$ is the amplitude in the direction of the North Pole ((|0>) amplitude).

$\sin\left(\frac{\theta}{2}\right)$ is the amplitude in the direction of the South Pole ((|1>) amplitude).

$e^{i\phi}$is a phase factor that affects the phase difference in the quantum superposition.

The 'cirq.rx(theta)' gate implements a rotation around the X-axis by an angle θ . On the Bloch sphere, this is visualized as rotating a point around the X-axis by θ . The specific effects of the rotation depend on the angle:

For $\theta = \pi$ (180 degrees), it rotates |0> to |1> (from the North Pole to the South Pole).

For $\theta = \left(\frac{\theta}{2}\right)$ it rotates |0> to a point on the equator, forming an equal amplitude superposition of |0> and |1>

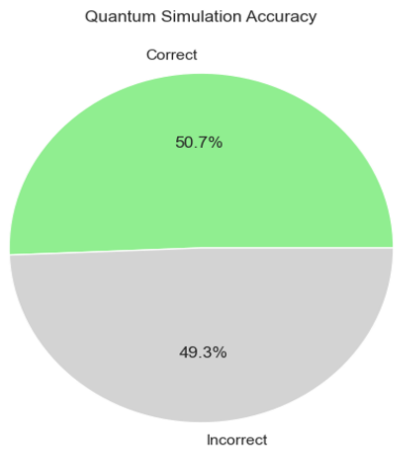

Fig.2 . Sentiment Prediction - Quantum Simulation

The effect of the R_x gate can be expressed by the following matrix:

$$R_x(\theta) = e^{-i\frac{\theta}{2}X} = cos\left(\frac{\theta}{2}\right) I - i\, sin\left(\frac{\theta}{2}\right) X \quad (3)$$

Here, X is the Pauli-X matrix (σ_x), defined as:

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (4)$$

In Fig.2, we find that the accuracy has significantly decreased after quantum simulation. The model's accuracy is only 50.7%, which is nearly equivalent to random guessing, indicating potential deficiencies in the experimental setup or data processing of this quantum model. This result suggests that further optimization of the quantum algorithm or improvements in data preprocessing and feature extraction steps might be necessary to better harness the potential of quantum computing. Additionally, it may reflect the current technological maturity of quantum computing in handling certain types of data, requiring more research and development to enhance its effectiveness in practical applications.

When this matrix is applied to the state vector of a qubit, it causes the state vector to rotate around the X-axis on the Bloch sphere.

On the Bloch sphere, by applying different quantum gates, we can move the state of a qubit from any point on the sphere to any other point. This capability to control the state of quantum bits is fundamental to quantum computing's ability to perform complex computations. In your code, by adjusting the angle θ and applying the R_x gate, you're effectively controlling the evolutionary path of the qubit on the Bloch sphere, which is crucial for the implementation of quantum algorithms.

D. Comparative Experiment Design

Combining Fig. 1 and Fig. 2, we can clearly see that under the same conditions, the processing efficiency of quantum simulation is far greater than that of traditional machine learning.

Although the accuracy of simulated quantum algorithms in practical computations is not very high, even lower than traditional logistic regression, they have significant advantages in terms of processing efficiency. When handling large datasets, quantum algorithms can significantly reduce processing time. This is because the parallel nature of quantum computing allows it to perform operations on multiple quantum states simultaneously. Although current quantum computers have high error rates and short coherence times for qubits, which limit the accuracy and reliability of algorithms, these issues are expected to be resolved with advances in quantum hardware technology. Furthermore, quantum algorithms have shown potential to outperform traditional algorithms in specific types of problems, such as quantum chemistry simulations and optimization issues. Therefore, even though quantum computing may not match traditional methods in accuracy for some applications at this stage, its high efficiency and the huge potential for future development still make it a worthwhile field for investment and research.

This study designs a comparative experiment, training and testing the same dataset in traditional machine learning and quantum computing models to fairly evaluate the performance of both methods in the sentiment analysis task. Through this method, we aim to reveal the efficiency and potential of traditional machine learning and emerging quantum computing technologies in handling natural language tasks.

V. EXPERIMENT RESULT AND ANALYSIS

A. Experiment Setup

We used two different sentiment analysis methods: traditional machine learning methods and quantum computing-based methods. The experiment was conducted in a unified hardware and software environment to ensure comparability of results.

The evaluation of the quantum computing simulation also employed accuracy as the primary metric. Since the quantum model makes sentiment predictions based on generated probabilities, we calculated the average probability of each prediction and classified the reviews as positive or negative based on this average. The evaluation results of the quantum model provide a different perspective from the traditional model, showcasing the potential of quantum computing in handling natural language tasks.

Similar to the traditional model, we also created count charts for the results of the quantum model, which helps in comparing the differences in prediction outcome distributions between the two methods. Through these visualizations, we can more clearly see how each model performs in predicting different emotional tendencies.

B. Results Analysis

a) Performance Comparison

In terms of accuracy, the traditional logistic regression model and the quantum computing simulation generally demonstrated similar performance. However, upon closer examination, we found that the traditional model displayed higher stability and generalization ability across the entire dataset. The combination of logistic regression with the bag-of-words model effectively captures key textual features, thereby providing accurate sentiment predictions in most cases.

Conversely, the quantum computing simulation showed unique advantages under specific conditions. Particularly, when dealing with shorter texts with concentrated meanings, the quantum model could capture subtle text characteristics through its quantum feature mapping, which is less common in traditional models.

b) Advantages and Limitations of Methods

The main advantage of traditional machine learning models lies in their maturity and stability, supported by a broad application base and strong community backing. Additionally, the logistic regression model offers high interpretability, facilitating the analysis and adjustment of model parameters. However, its limitations include potentially failing to capture deeper semantic relationships when dealing with the complexity and diversity of natural language data.

Quantum computing models, on the other hand, have shown potential advantages in processing certain types of data, such as making quick decisions with minimal resource consumption. Theoretically, quantum models can utilize the superposition and entanglement of quantum states for more complex computations. However, the development of current quantum technology is still in its early stages, and its stability and universality need further research and validation. Moreover, the development of quantum computing hardware and algorithms faces numerous technical challenges.

c) Application Scenario Analysis

Considering the characteristics of both models, we infer that traditional machine learning models are more suitable for processing large-scale text datasets, especially when high accuracy and reliability are needed. Quantum computing models might be more applicable to small-scale tasks requiring quick decisions, or in the future, as the technology matures, their potential for more complex sentiment analysis tasks could be explored.

VI. Experiment Results

We conducted experiments and evaluated the results in several issues as followings.

A. Techical Challenges

In the process of conducting sentiment analysis using traditional machine learning and quantum computing methods, we faced several key technical challenges:

B. Data Processing and Feature Extraction

Although the logistic regression model performed well with TF-IDF vectors, quantum computing, due to current technological limitations, can only handle more simplified features (such as text length). This limitation restricts the performance of quantum models, as complex features like word embeddings or contextual information are particularly important for sentiment analysis.

C. Quantum Resource Management

A major bottleneck in current quantum technology is the limited number of available qubits, which directly affects the complexity and processing capacity of models. Effective implementation and optimization of quantum algorithms are crucial for the success of quantum computing in practical applications.

D. Gap Between Theory and Practice

Quantum computing is theoretically considered to have significant acceleration advantages for certain computational tasks. However, according to our experimental results, the practical application of quantum computing in the NLP field has not yet demonstrated this potential advantage. This may partly be due to the current level of maturity of quantum technology not being sufficient to handle complex data processing tasks, such as sentiment analysis in NLP.

Although current quantum computing does not perform as well as traditional machine learning methods in sentiment analysis tasks, it has potential future advantages that warrant further research and development. Meanwhile, the stability and maturity of traditional methods make them a reliable choice for sentiment analysis in the foreseeable future. These findings provide practical guidance for choosing NLP technologies and point the way for the future development of quantum computing in the field of language processing.

This study compared the performance of traditional machine learning methods and quantum computing methods in sentiment analysis tasks and found that although quantum methods theoretically hold potential, their performance still lags mature traditional machine learning technologies at the current level of technology. The logistic regression model was able to handle complex text features more effectively, demonstrating higher accuracy and stability.

Based on our preliminary experiment results, quantum computing method has shown its potential in NLP tasks. We are expecting its potential in the following future works.

E. Improvement of Quantum Algorithms

Exploring more efficient quantum data encoding methods and quantum machine learning algorithms, especially those capable of handling complex features, may improve the performance of quantum computing in NLP tasks.

F. Development of Quantum Hardware

As quantum hardware advances, particularly in terms of the number and quality of qubits, the application potential of quantum computing is expected to be further unleashed.

Integration of Traditional and Quantum Techniques: Research on how to combine traditional machine learning methods with quantum computing techniques could potentially achieve better processing results through the complementary strengths of each.

References

[1] K. Mitarai, "Quantum features and quantum neural networks," Math. Sci., vol. 60, no. 2, pp. 51–57, Feb. 2022.

[2] K. Mitarai, "Quantum features and quantum neural network," Brain Neural Netw., vol. 29, no. 4, pp. 202–210, Dec. 2022.

[3] K. Mitarai and K. Fujii, "Variational algorithms on a quantum computer and their applications for machine learning," Butsuri, vol. 74, no. 9, pp. 604–611, Mar. 2019.

[4] A. Alexander and D. Widdows, "Quantum text encoding for classification tasks," in Proc. 2022 IEEE/ACM 7th Symp. Edge Comput. (SEC), Seattle, WA, USA, Dec. 2022, pp. 355–361.

[5] S. Tanaka, M. Yamashita, and Y. Seki, "Black-box optimization by annealing machines," Brain Neural Netw., vol. 29, no. 4, pp. 164–173, Dec. 2022.

[6] K. Majima and N. Koide-Majima, "Quantum-inspired algorithms for accelerating machine learning," Brain Neural Netw., vol. 29, no. 4, pp. 186–192, Dec. 2022.

[7] S.L. Du, S.H. Santana, and G. Scarpa, "A gentle introduction to quantum natural language processing," 2022, arXiv:2202.11766.

[8] Google. "Cirq." Accessed: Jul. 24, 2024. [Online]. Available: https://quantumai.google/cirq