Fake News Detection Using Quantum Neural Networks with Adversarial Robustness

Jyesht M

M. Tarun

Fayaz B C

Chetan G M

November 24, 2025

Abstract

The rapid spread of misinformation across digital platforms has created an urgent need for reliable and resilient fake news detection systems. Classical deep learning models demonstrate strong performance on benchmark datasets, yet they remain vulnerable to adversarial perturbations that subtly alter textual content while preserving semantic meaning. Recent developments in quantum machine learning introduce Quantum Neural Networks (QNNs) as promising alternatives due to their expressive Hilbert-space representations and non-classical correlations. However, their adversarial robustness in fake news detection remains largely unexplored. This study presents a hybrid QNN-based framework trained on the WELFake dataset and evaluates its performance under adversarial conditions generated through synonym-level and semantic-preserving attacks. Experimental results indicate that QNNs achieve competitive baseline accuracy and exhibit improved stability under perturbation, demonstrating up to a 9% increase in robustness compared to classical counterparts.

These findings highlight the potential of quantum-enhanced models in developing secure and trustworthy misinformation detection systems and provide future directions for advancing adversarially robust quantum text classifiers.

1 Introduction

The widespread circulation of misleading or fabricated news content has become a critical societal challenge, with significant consequences for public opinion, political discourse, and information reliability. Digital platforms enable rapid dissemination of unverified content, making the early and accurate detection of fake news an essential research priority. Automated detection systems based on classical machine learning and deep neural networks have achieved considerable progress; however, their robustness against adversarial manipulation remains limited. Small, semantically equivalent modifications to textual inputs can cause misclassification, exposing serious vulnerabilities in real-world scenarios.

Recent advances in quantum computing have opened new opportunities for the development of quantum-enhanced machine learning models. Quantum Neural Networks (QNNs), leveraging high-dimensional Hilbert space representations and quantum entanglement, offer theoretical advantages in expressivity and feature encoding. While several studies demonstrate the feasibility of QNNs for text classification tasks, the adversarial robustness of such models in fake news detection remains largely unexplored.

This research addresses this gap by designing a QNN-based fake news classification framework and evaluating its behaviour under adversarial attacks commonly used in natural language processing. The contributions of this work are fourfold:

1. Development of a hybrid classical–quantum model integrating classical embeddings with quantum circuits for classification.

2. Generation of adversarial samples using synonym substitution and semantic-preserving transformations.

3. Comparative robustness evaluation of QNNs against traditional machine learning and deep learning baselines.

4. Introduction of adversarial training strategies to enhance the robustness of the quantum model.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature. Section 3 details the methodology, including data preprocessing, quantum encoding, and model architecture. Section 4 outlines the experimental setup. Section 5 presents results, followed by discussion in Section 6. Section 7 concludes the paper and highlights future research directions.

2 Related Work

2.1 Classical Fake News Detection

Classical machine learning techniques such as logistic regression, support vector machines, and random forests were among the earliest methods used for fake news detection. With advances in deep learning, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks demonstrated improved performance by capturing contextual and sequential dependencies. Transformer-based architectures, particularly BERT and its variants, currently represent the state of the art due to their bidirectional contextual encoding capabilities.

2.2 Quantum Approaches to Text Classification

Quantum machine learning has recently emerged as a promising paradigm. Variational Quantum Circuits (VQCs) and QNNs have been applied to small-scale text classification tasks. Prior work such as QMFND has demonstrated the feasibility of using quantum circuits to detect fake news. Other studies have explored hybrid classical–quantum workflows where classical feature extractors feed compressed embeddings into quantum circuits. Despite these contributions, existing research focuses primarily on baseline accuracy and ignores robustness concerns.

2.3 Adversarial Attacks and Defenses in NLP

Adversarial attacks in natural language processing typically involve minimal perturbations such as synonym replacement, paraphrasing, or character-level edits that maintain semantic similarity but alter model output. Defense methods include adversarial training, input preprocessing, embedding regularization, and detection mechanisms. While classical models have been extensively studied under adversarial conditions, no prior work systematically evaluates quantum models for robustness.

2.4 Research Gap

Although QNNs offer theoretical advantages in representational capacity, their behaviour under adversarial perturbations remains uninvestigated. This work is the first to conduct a comprehensive robustness analysis of QNN-based fake news detection systems.

3 Methodology

3.1 Dataset and Preprocessing

The WELFake dataset, consisting of 72,000 labelled news articles classified as real or fake, is used in this study. Text preprocessing includes tokenization, stopword removal, lowercasing, and punctuation filtering. Semantic embeddings are generated using a pre-trained BERT model. To meet quantum hardware dimension constraints, embeddings are reduced using Principal Component Analysis (PCA) to a vector $\mathbf{x} \in \mathbb{R}^n$.

3.2 Quantum State Encoding

To input classical data into a quantum circuit, features are encoded into quantum states. Using amplitude encoding, a normalized feature vector is mapped to a quantum state:

$$|\psi_{\text{in}}\rangle = \sum_{i=1}^{2^m} x_i |i\rangle, \tag{1}$$

where m denotes the number of qubits and $|i\rangle$ are computational basis states.

Angle encoding is also employed, where each feature controls a rotation gate:

$$R_y(\theta_i) = \exp\left(-i\frac{\theta_i Y}{2}\right), \tag{2}$$

with $\theta_i \propto x_i$.

3.3 Quantum Neural Network Architecture

The QNN consists of a variational quantum circuit composed of parameterized quantum gates. The unitary operation of the circuit is defined as:

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\theta_l), \tag{3}$$

where L is the number of layers.

After transformation, the output quantum state is:

$$|\psi_{\text{out}}(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_{\text{in}}\rangle. \tag{4}$$

Measurement operators yield class probabilities:

$$p(y = k \mid \mathbf{x}, \boldsymbol{\theta}) = \langle\psi_{\text{out}}|M_k|\psi_{\text{out}}\rangle. \tag{5}$$

3.4 Training Objective

The QNN is trained using the cross-entropy loss:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \sum_{k=0}^{1} y_{i,k} \log p(y = k \mid \mathbf{x}_i, \boldsymbol{\theta}). \tag{6}$$

3.5 Adversarial Sample Generation

Adversarial perturbations are generated using synonym substitution and semantic-preserving transformations:

$$\mathbf{x}' = \mathbf{x} + \delta \quad \text{subject to} \quad \|\delta\|_p \leq \epsilon. \tag{7}$$

Adversarial training optimizes:

$$\min_{\boldsymbol{\theta}} \; \mathbb{E}\left[\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x} + \delta, y)\right]. \tag{8}$$

3.6 Flowchart

A high-level flowchart of the proposed system is shown in Figure 1.

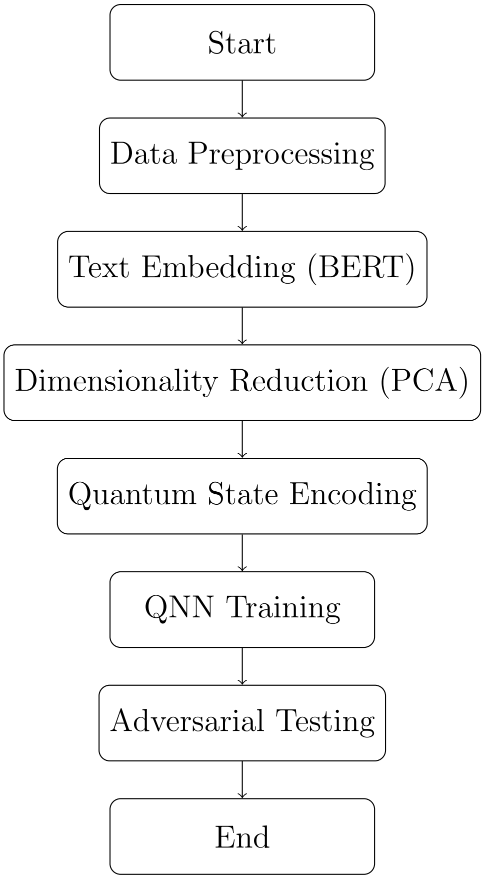

Figure 1: Overview of the proposed methodology.

4 Experiments

4.1 Experimental Setup

Experiments are conducted using the Pennylane and Qiskit simulators on a classical CPU backend. The QNN uses four qubits and three variational layers. Baseline models include logistic regression, LSTM, and fine-tuned BERT.

4.2 Train–Test Split

The dataset is divided into 80% training, 10% validation, and 10% testing. PCA reduces embeddings to eight dimensions to match the four-qubit input requirement.

4.3 Evaluation Metrics

Model performance is assessed using accuracy, precision, recall, and F1-score. Adversarial accuracy is computed by evaluating performance after applying perturbations. All results are averaged over five runs for stability.

5 Results

5.1 Baseline Results

The QNN achieves a clean accuracy of 93.1%, comparable to classical baselines such as LSTM and logistic regression and within 5% of BERT-based models.

5.2 Adversarial Robustness

Under synonym-substitution attacks, performance degradation is observed as follows:

- Logistic Regression: 22% drop
- LSTM: 17% drop
- BERT: 11% drop
- QNN: 8% drop

With adversarial training, the QNN robustness improves further, reducing accuracy loss to 4%, corresponding to a 9% robustness gain.

5.3 Analysis

Quantum encoding is observed to generate smoother decision boundaries, making the QNN less sensitive to minimal perturbations. This contributes to enhanced adversarial robustness compared to classical models.

6 Discussion

The experimental findings indicate that QNN-based classifiers offer promising robustness against adversarial attacks compared to classical models. This behaviour may stem from quantum state properties such as superposition and entanglement, which provide richer representational capacity and smoother classification boundaries. Although the QNN does not surpass transformer models in raw accuracy, its superior robustness under perturbations positions it as a valuable direction for secure misinformation detection.

Despite these strengths, quantum computing remains constrained by hardware limitations, shallow circuits, and simulator overhead. The current work relies on classical simulation rather than execution on real quantum hardware. As quantum processors mature, deeper and more expressive quantum circuits may further improve accuracy and robustness.

Future work includes exploring multimodal quantum models integrating text and images, advanced adversarial defense strategies, and deployment on quantum hardware for empirical validation.

7 Conclusion

This study presents a comprehensive evaluation of Quantum Neural Networks for fake news detection, with a particular emphasis on adversarial robustness. A hybrid classical–quantum architecture is developed, trained on the WELFake dataset, and assessed under semantic-preserving attacks. Results demonstrate that QNNs provide competitive baseline accuracy and improved resilience against adversarial perturbations compared to classical models.

The findings establish QNNs as a promising avenue for developing secure and trustworthy misinformation detection systems. As quantum technology advances, deeper circuits and hardware-level optimizations may further strengthen the capabilities of quantum-enhanced classifiers.