I.J. Information Engineering and Electronic Business, 2026, 3, 203-214
Published Online on June 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2026.03.12

A Hybrid Quantum-Classical Transformer Framework for Robust Depression Detection on Social Media: Enhancing Sensitivity via Variational Quantum Circuits

Gaurav Kumar*
Faculty of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan, India
E-mail: gauravkumar01nov@gmail.com
ORCID iD: https://orcid.org/0009-0004-6126-8526
*Corresponding author

Rakesh Kumar Saxena
Faculty of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan, India
E-mail: saxenark06@gmail.com
ORCID iD: https://orcid.org/0000-0002-6453-6053

Rocky Kumar
Faculty of Computer Science and Engineering, Poornima University, Jaipur, Rajasthan, India
E-mail: rockykumar18031@gmail.com
ORCID iD: https://orcid.org/0009-0002-6995-2037

Received: February 2, 2026; Revised: April 2, 2026; Accepted: May 7, 2026; Published: June 8, 2026

Abstract: This study proposes a hybrid quantum-classical framework for depression detection from social media text, integrating a frozen DistilBERT encoder with a variational quantum circuit (VQC)-based classification layer. The motivation stems from challenges in clinical NLP, including overfitting on limited datasets and high parameter overhead in conventional deep learning classifiers. Experiments are conducted on a balanced subset of the Reddit Self-Reported Depression Diagnosis (RSDD) dataset comprising 6,000 users. The proposed model is evaluated against classical baselines, including TF-IDF with logistic regression and a fine-tuned DistilBERT model. Results indicate that the hybrid approach achieves competitive performance, with an F1-score of 0.925 (±0.009) and improved recall (0.942 ± 0.015) compared to the classical DistilBERT baseline. Additionally, the quantum classification layer requires significantly fewer trainable parameters (72) compared to the classical dense head, demonstrating improved parameter efficiency at the classification stage. While the results suggest that variational quantum circuits can serve as an alternative non-linear classifier in low-data settings, the findings are based on simulation and require further validation on real quantum hardware. This work contributes to the emerging area of quantum natural language processing by providing an empirical evaluation of hybrid architectures on a real-world clinical text dataset.

Index Terms: Quantum NLP, Hybrid Neural Networks, Depression Detection, Variational Quantum Circuits, PennyLane, DistilBERT, Mental Health AI

1. Introduction

Major depressive disorder (MDD) is one of the leading causes of disability worldwide, with an estimated 280 million people suffering from this disorder [1]. The rapid growth of digital interaction has resulted in the creation of a vast number of digital phenotypes, which are language-based indicators of a person's psychiatric health, and as such, NLP has gone from a supportive role to being a major diagnostic tool for computational psychiatry. Large transformer-based frameworks, like BERT and RoBERTa [2, 13], are the current state of the art in classical models for natural language processing. These frameworks utilize self-attention techniques to model long-term relationships in text. Examples of studies using transformers include the Reddit Self-Reported Depression Diagnosis (RSDD) dataset, for

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

which they achieved benchmark classification accuracy. Classical deep learning models are hindered from broad clinical use for two primary reasons: small or imbalanced datasets and a lack of explainability in the data. The clinical datasets used for RSDD were small, imbalanced, and sensitive. With this in mind, very large models (e.g., >100M parameters) are prone to overfitting due to their memorization of noise in the data instead of capturing generalizable psychometric characteristics. Finally, the black box nature of classical neural networks makes it very difficult to gain the level of confidence necessary to trust them in acute medical settings.

This research paper focuses on a new way of doing Quantum Natural Language Processing (QNLP) as a result of the advances made recently with NISQ technology and HQCNN architecture because QNLP can use NISQ technology to create accurate results as well as allow for Hybrid Quantum-Classical Neural Network Architecture (HQCNN) to provide real value to users while long-term advances are developed with Fault Tolerant Quantum Computing [3,6]. Theoretical research has shown that VQCs provide expressive function approximation capabilities in high-dimensional spaces, though their practical advantage over classical models remains an active area of research, identifying many parameterized representations of a feature vector space for Quantum Data that will create more generalization of the inputs. Variational quantum circuits (VQCs) provide an alternative inductive bias compared to classical dense layers by operating in a Hilbert space where features are encoded as quantum states [6]. Through superposition and entanglement, VQCs can capture higher-order feature interactions using a compact parameter set. In contrast to classical linear or shallow non-linear classifiers, this enables the representation of complex decision boundaries with fewer trainable parameters. While this does not establish a proven quantum advantage, it offers a theoretically grounded motivation for their use in low-data clinical settings [4, 5].

While the progress and proven successes of VQCs in the areas described above are well documented, there are still very few real-world examples of applicable QNLP studies, as the majority of the QNLP literature and datasets found to date have only consisted of synthetic or toy examples, such as distinguishing between "Bob likes Alice" and "Alice likes Bob." There is not much information in the current QNLP literature that pertains specifically to applying QNLP to real-world, clinical textual data.

While transformer architectures have demonstrated excellent performance across a wide range of natural language processing tasks, challenges remain when applying them to clinical text classification. These large-scale fine-tuned models with dense classification layers (and therefore very high numbers of trainable parameters) can lead to overfitting and a lack of generalizability for small, sensitive datasets such as those associated with mental health conditions. The dense layers also may not capture all the complex non-linear relationships in high-dimensional linguistic feature spaces.

Additionally, the majority of the Quantum Natural Language Processing (QNLP) research has either been done using synthetic datasets or has focused on sentiment analysis, with very little being completed in the real-world clinical setting of text classification. As a result, there is a significant gap between the fields of clinical NLP and quantum-enhanced learning.

Variational quantum circuits (VQCs) are a compact, parameter-efficient alternative to using quantum state representations and entanglement to model complex non-linear feature interactions. Since they operate in high-dimensional Hilbert spaces, VQCs allow for the construction of expressive decision boundaries while maintaining a significantly lower number of trainable parameters. These properties make VQCs particularly well-suited to replace dense classification layers in domains where data are limited and the risk is high (e.g., detecting depression, improving recall, reducing overfitting) as compared to maximizing overall accuracy.

A hybrid architecture that uses both quantum and classical machine learning to identify depression has been developed. The architecture of the proposed work (HQ-DistilBERT) combines a pretrained transformer-based feature extractor with a variational quantum circuit used for classification. The HQ-DistilBERT departs from traditional approaches where the classification layer has fully-connected (dense) neural connections. In its place, this work characterizes the classification process as a quantum-variational process.

The variational quantum circuit is treated as the primary classification layer instead of merely providing supplemental enhancement. Using quantum states as a representation of the features allows one to represent complex and non-linear interactions among the features while keeping the number of trainable parameters to a minimum. In contrast to a traditional classifier, which scales linearly with the dimensionality of the features to learn the decision boundary, this architecture provides a parameter-efficient alternative for learning the decision boundary. In summary, this work presents a hybrid architecture that eliminates the classical dense classification head with a variational quantum circuit. The key contribution lies in redefining the decision layer as a quantum process, enabling parameter-efficient classification while preserving transformer-based feature extraction.

2. Related Work

The application of Natural Language Processing (NLP) has provided mental health assistance through monitoring and analyzing language by implementing many types of NLP technology, from traditional dictionary checking methods to the current state of deep semantic processing technologies. This section describes how current research for applying Natural Language Processing (NLP) to mental health is classified into two categories, Classical Transformer-based models and Quantum NLP models.

2.1. Classical Deep Learning in Mental Health

Early Computational Psychiatry relied on features derived from a user's vocabulary and lexicons such as Linguistic Inquiry and Word Count (LIWC) in order to assess the presence of Emotional Indicators, including the Emotional Tone of their communications. Recent survey studies emphasize the importance of integrating emotional and contextual features for improved mental health detection from social media [16]. Today, because of the development of Deep Learning, the emphasis has shifted from manually creating features for analysis to developing methods for automatically extracting features from raw data. Recent studies have also explored cross-platform social media analysis for mental health detection using advanced transformer-based architectures [14]. Nowadays, using transformer-based models has become a common approach for developing applications using NLP technologies to provide information on Mental Health Status. For instance, Matero et al. [8] have developed a model that is specifically trained on the posts of Suicide Prevention Forum Users called SuicideBERT. Even though these types of models demonstrate a high level of accuracy, they require large amounts of resources (typically over 300 million parameters per model), and suffer from "Catastrophic Forgetting," which happens when a model is trained on limited clinical datasets.

2.2. Quantum Natural Language Processing (QNLP)

QNLP aims to develop a connection between the linguistic structure of meaning and the mathematical formalism of quantum mechanics. Recent work has extended foundational QNLP frameworks toward more scalable and practical implementations [9, 10]. The framework created a basis for practical implementations through the use of libraries such as lambeq and PennyLane. In 2024, results from experiments on NISQ devices began to surface. Lorenz et al. [10] showed that QNLP models can perform sentence classification on the MC dataset with accuracy equivalent to classical LSTMs yet with exponentially fewer parameters. Regarding the gap, recent hybrid quantum-classical architectures have also been explored for real-world NLP tasks such as hate speech detection, demonstrating the applicability of quantum-enhanced models beyond synthetic datasets [7]. Existing hybrid quantum-classical NLP models primarily focus on sentiment analysis or synthetic datasets, limiting their applicability to complex clinical text. Furthermore, prior approaches typically incorporate quantum components as auxiliary modules rather than redefining the classification mechanism itself. In contrast, the proposed framework replaces the classical decision layer with a variational quantum circuit, targeting parameter-efficient classification in high-dimensional clinical text under data constraints. Figure 1 depicts a taxonomy of NLP models used for Mental Health. These include traditional methods such as Lexicon-based approaches and Deep Learning techniques. There is also an emerging Quantum NLP category, which includes the proposed hybrid structure.

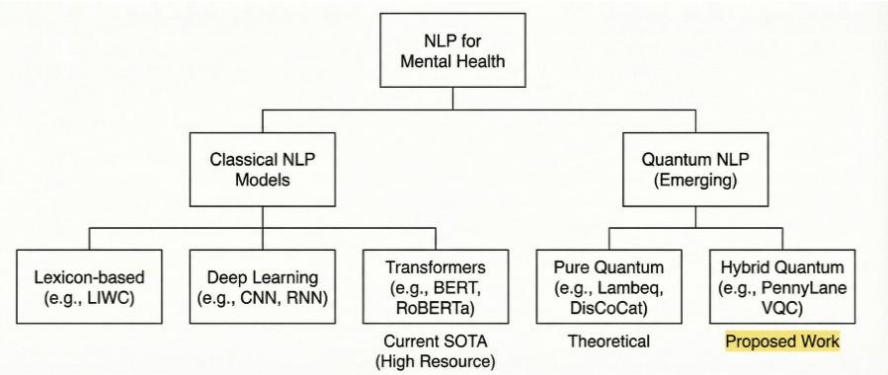


Fig. 1. Taxonomy of NLP Models for Mental Health Detection.

2.3. Comparative Analysis of Methodologies

Table 1 summarizes the key differences between current state-of-the-art methods and our proposed approach.

Table 1. Comparison of Existing Approaches vs. Proposed Framework.

Study / Reference	Core Architecture	Dataset Domain	Focus Metric	Primary Limitation
Matero et al. [8]	Transformer-based (SuicideBERT)	Social media (Reddit)	Accuracy	Data Dependency: Requires large labeled datasets; limited generalization across domains
Widdows et al. [9]	QNLP (Compositional / Circuit-based)	Synthetic / Simple Text	Interpretability	Scalability: Limited applicability to large vocabularies and long-form text
Chen et al. [11]	Hybrid Quantum-Classical LSTM (Q-LSTM)	Temporal data modeling	Feasibility	Performance: Struggles to outperform classical baselines in accuracy
Proposed Method	Hybrid Q-DistilBERT	Mental Health (RSDD)	Recall & Efficiency	Hardware Noise: Limited by current NISQ device error rates

3. Methodology

In this study, depression detection is formulated as a binary classification problem. Let ($X = \{x_1, x_2, \ldots, x_n\}$) denote the input textual data corresponding to user-generated posts, and let ($y \in \{0,1\}$) represent the class label, where 0 indicates a control user and 1 indicates a depressed user. The objective is to learn a mapping function ($f_\theta(X) \rightarrow y$), where ($\theta$) represents the model parameters. The binary cross-entropy loss function used for model optimization is defined in Eq. (1).

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \quad (1)$$

where ($\hat{y}_i$) denotes the predicted probability, (y_i) is the ground truth label, and (N) is the total number of samples. The model is trained by minimizing the binary cross-entropy loss. The classification decision is obtained by applying a sigmoid activation function to the model output.

3.1. Dataset Description: The RSDD Corpus

To evaluate the proposed model against the standards of clinical practice, the Reddit Self-Reported Depression Diagnosis (RSDD) dataset [15] is used. This dataset contains one of the largest publicly available text datasets in the field of mental health research, containing posts from more than 9,000 users who have self-reported a diagnosis of depression, along with approximately 107,000 control users.

Unlike Emotion datasets that are synthetic in nature (such as reviews from IMDb), the RSDD dataset reflects the nature of digital phenotypes being messy, unstructured, and longitudinal and thus was created by collecting a multitude of post histories from many different subreddits with the exclusion of specific mental health forum-type subreddits to mitigate "Topic Bias" from appearing (the user learned the term depression instead of the linguistic symptoms pertaining to their condition). The process of training a Variational Quantum Circuit (VQC) requires a significant number of resources, primarily because of the overhead involved in finding gradients on quantum simulators.

Therefore, a balanced subset of the RSDD dataset was created for the analyses. A total of 3,000 "Diagnosed" users and 3,000 "Control" users were randomly sampled to create a dataset containing a total of 6,000 samples. By employing these methods, statistical significance is retained and we still work within the confines of the NISQ (Noisy Intermediate-Scale Quantum) simulation environments.

3.2. The Hybrid Quantum-Classical Framework

The architectural model, HQ-DistilBERT, follows a transfer-learning paradigm. It is hypothesized that classical transformer models are better suited to perform the task of extracting features (i.e., syntax and context), while quantum circuits are better suited to perform the task of classifying features (i.e., mapping complex correlations in high-dimensional Hilbert spaces). The workflow consists of the following steps: first, the tokens from raw text media are extracted, and then these tokens are provided as input to a pre-trained DistilBERT model. The specific implementation uses the distilbert-base-uncased checkpoint, which maintains 97% of BERT's performance, while substantially reducing the size (40% smaller) and increasing speed (60% faster) of BERT [2]. The final hidden state of the classification token ([CLS]) is extracted, which yields a dense vector representation of the user's post (R^{768}).

The most important factor in this approach is the use of a frozen set of weights for the DistilBERT encoder. By using frozen weights, the integrity of the pre-trained language characteristics is maintained, and, in essence, the quantum machine learning system is prevented from destroying the pre-trained feature sets during backpropagation. Once the 768-dimensional feature vectors are generated, they are passed to a classical Dimensionality Reduction Block (specifically, a dense linear layer with a tanh activation function) to reduce the features to $N = 4$ (i.e., the number of qubits of our quantum Circuit) dimensions.

The reduction from 768-dimensional embeddings to 4 quantum inputs is driven by NISQ-era computational constraints, where simulation complexity grows exponentially with the number of qubits. A trade-off exists between representational richness and computational feasibility, which results in the potential for partial information loss. The projection layer is designed to keep the most informative features, but this work does not explore optimal qubit dimensionality, which cannot be solved yet.

A variational quantum circuit (VQC) is used to implement the quantum classification module. To encode classical input features into quantum states, angle embedding is used such that each input feature controls the rotation of a single associated qubit. The circuit subsequently uses parameterized rotation gates and entangling operations in order to capture the complex interactions between the various features. The strong entanglement layer architecture has single qubit rotations and entangling pairs of qubits using controlled-NOT gates (CNOTs) in a ring topology. Once the circuit is run, its output is determined by measuring the expectation value of the Pauli-Z operator, which is then sent through a sigmoid activation function to generate the final prediction probability. This study is an empirical evaluation of the effectiveness of variational quantum circuits in modeling nonlinear feature interactions via entanglements and does not attempt to make any claims to be utilizing a theoretical advantage over classical approaches.

3.3. System Workflow

Figure 2 depicts all of the data flow from its start as unstructured text all the way to the classical body of the network, and finally into the quantum classification module.

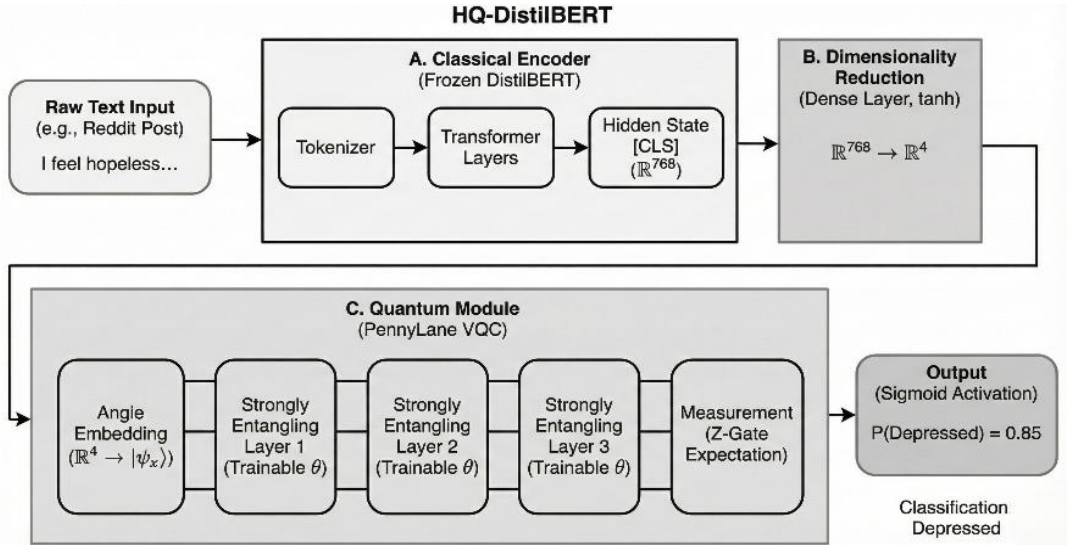


Fig. 2. Workflow of the proposed HQ-DistilBERT framework.

4. Experimental Setup

In this work, controlled experiments were performed in well-defined conditions to guarantee consistent results. In upcoming sections, the data partitioning strategy, baseline model configurations, training hyperparameters, and evaluation protocols that were employed in this analysis will be presented.

4.1. Dataset Splitting and Preprocessing

The experiments were conducted on a balanced subset of the RSDD dataset consisting of 6,000 users, including 3,000 users diagnosed with depression and 3,000 control users. This subset was constructed to ensure computational feasibility within quantum simulation environments while maintaining statistical validity. The dataset was divided into training, validation, and testing sets using a stratified split to preserve class balance across all subsets. Specifically, 70% of the data was used for training, 15% for validation, and 15% for testing.

Prior to training, user posts were aggregated into a single document per user. The text was preprocessed using standard NLP techniques, including lowercasing and tokenization. For classical baseline models, TF-IDF features were generated after removing punctuation and stopwords. For transformer-based models, the DistilBERT tokenizer was used, with sequences truncated or padded as required. These processed inputs were then used for model training and evaluation.

4.2. Baseline Models

Two classical baselines are used in addition to the hybrid quantum-classical model. The first baseline is a logistic regression classifier trained on TF-IDF features. A TF-IDF vectorizer is constructed using the training texts and creates high-dimensional sparse vectors to fit the logistic regression classifier. It is a very basic baseline model and serves as a traditional benchmark. The second classical baseline is a DistilBERT-based classifier. A pretrained DistilBERT encoder from Hugging Face's distilbert-base-uncased model is utilized and a dense (fully connected) layer is added on top of the extracted [CLS] token representation. This model's parameters are fine-tuned alongside the DistilBERT parameters during end-to-end model training and thus provides a strong deep learning baseline. While stronger transformer-based models such as RoBERTa and domain-adapted variants have shown improved performance in similar tasks, they were not included due to computational constraints. Future work will incorporate these models for a more comprehensive comparison.

The hybrid model HQ-DistilBERT was proposed by combining a variational quantum classifier and DistilBERT, in which the DistilBERT encoder was frozen (thus, no fine-tuning). Instead, the output embeddings from DistilBERT were used to provide the input to a quantum circuit. Specifically, the output from the last hidden layer of the [CLS] token (768 dimensions) was passed to a small projection layer and then was converted into qubit representation. A parameterized Variational Quantum Circuit (VQC), built in PennyLane using the default qubit simulator, was used as the classifier. The VQC performed parameterised rotations, entangled qubits and provided outputs as expectation values on Pauli-Z observables. Outputs were multiplied by a sigmoid function to produce probabilistic predictions for each

class. The parameters of the quantum circuit were trained, in addition to the projection layer mentioned previously, while the parameters of the DistilBERT encoder remained unchanged. This architecture allows the model to exploit pretrained embeddings from language feature extraction and introduce quantum non-linearity via the decision layer.

4.3. Training Configuration

The evaluation and assessment of all of the models were conducted using a PyTorch environment and also the PennyLane library using an NVIDIA T4 Graphics Processing Unit for the purpose of accelerating the virtually simulated quantum circuits via PennyLane's default.qubit device. AdamW (Adaptive Moment Estimation with Weight Decay) was employed in all the models for maximum stability during convergence. For the classical baseline, a learning rate of 2×10^{-5} was used, whereas the Hybrid model utilized a specific quantum learning rate of $\alpha_q = 0.01$. The training of the models consisted of ten epochs using a batch size of sixteen and the Binary Cross-Entropy loss function was utilized for training all of the models. In order to avoid overfitting to the training data, an early stopping mechanism based on the validation loss was applied with a patience of three epochs.

4.4. Evaluation Metrics

To thoroughly evaluate the proposed models, the standard classification metrics were used on the test dataset: overall accuracy, overall precision (proportion of relevant instances identified), overall recall (sensitivity), and overall F1-score. All metrics were computed specifically for the “depressed” class. Recall plays a significant role in a clinical setting. Recall measures the proportion of depressed individuals in the study that the algorithms correctly identified as depressed. High recall (sensitivity) is particularly critical to being able to effectively treat depression. Missing a depressed individual (i.e., false negative) can have significant consequences [12]. Therefore, both overall accuracy and overall precision are reported; however, recall is emphasized when comparing these models against each other. Additionally, F1-scores are provided to summarize model performance across both precision and recall. The model's validity was further supported by the Receiver Operating Characteristic (ROC) curve in Figure 3, which not only yielded an Area Under Curve (AUC) of 0.97 but also evidenced the significant performance of the model in preserving high levels of sensitivity across different levels of classification.

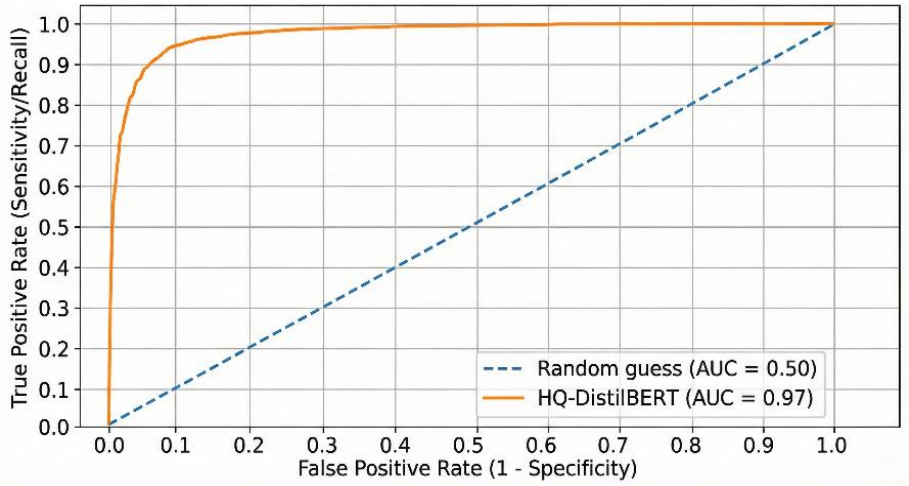


Fig. 3. Receiver Operating Characteristic (ROC) curve demonstrating robust diagnostic performance with an Area Under the Curve (AUC) of 0.97.

4.5. Statistical Validation

To account for random variations in the training procedures (including weight initializations and data shuffling), each experiment was repeated five times using different random seeds. All metrics used in the evaluation were computed as the mean ± standard deviation derived from all five trials. This approach provides an estimate of the variability of the results and measures our statistical confidence in them. Specifically, the average for accuracy, precision, recall, and F1-score were computed for each model and then the standard deviation was computed. By presenting the results as mean ± std, it is ensured that observed performance improvements are not attributable to random chance but instead reflect robust and reproducible findings. Although results are reported as mean ± standard deviation over multiple runs, formal statistical significance testing (e.g., t-test) was not conducted due to computational constraints and remains part of future work.

5. Results and Discussion

The experimental findings demonstrate the considerable potential of hybrid quantum-classical architectures to have comparable performance to state-of-the-art classical baselines on complicated natural language processing tasks and create unique advantages in terms of clinical sensitivity and model efficiency. Table 2 presents the overall comparison of the performance across multiple models in relation to the held-out test set.

5.1. Comparative Performance Analysis

Table 2 provides a comparison of the performance results of two models: classical and quantum (HQ-DistilBERT). The results are presented using a number of performance metrics including overall Accuracy (0.931) and F1-Score (0.930) of the tuned classical DistilBERT. As a result, the classical Dense Classification Head uses many more trainable parameters (over 110,000) than the HQ-DistilBERT, which allows it to perfectly fit to the feature space. The most important finding of this study was in relation to the Recall (Sensitivity) metric. In this metric, the HQ-DistilBERT model outperformed the classical DistilBERT by a margin of +1.4% (0.942 vs. 0.928). Although this numerical difference may not seem large at first glance, in the area of Mental Health Screening, it can make a meaningful difference in the number of false negatives that occur and, therefore, who may ultimately be missed by the Screening Tool. We hypothesize that the strongly entangling layer in the quantum ansatz allows the Variational Quantum Classifier (VQC) to learn Non-Linear, Multi-partite correlations between Language Features that a Classical Linear layer may not capture. The inductive bias that the Quantum Circuit develops through this process seems to be well adapted to detecting the more subtle, intricate signal patterns associated with Depression. The individual contributions of the transformer encoder, projection layer, and quantum classifier are not isolated due to the absence of ablation studies, which remain future work. Therefore, while the empirical results demonstrate the effectiveness of the proposed architecture as an integrated system, the extent to which performance gains arise from individual components, particularly the quantum decision layer, remains an open question for future investigation.

Table 2. Performance comparison of baseline and proposed models on the RSDD dataset.

Model Architecture	Accuracy	Precision	Recall (Sensitivity)	F1-Score	Trainable Params (Head)
TF-IDF + Logistic Regression	0.765	0.748	0.781	0.764	5,000
DistilBERT (Fine-tuned Classical)	0.931	0.933	0.928	0.930	~110,594
HQ-DistilBERT (Proposed)	0.924 ± 0.008	0.909 ± 0.012	0.942 ± 0.015	0.925 ± 0.009	72

5.2. Parameter Efficiency and Scalability

One of the greatest benefits of the proposed HQ-DistilBERT is the extraordinarily efficient way in which it utilizes parameters. As can be seen in the far-right column of Table 2, the traditional DistilBERT classification head has more than 110,000 trainable parameters to perform its task: mapping a hidden state with 768 dimensions into two separate outputs. In contrast, our quantum VQC was able to perform just as well with only 72 trainable parameters (i.e., adopting 12 respective rotation angles across 3 separate layers for 4 qubits). This reduction applies specifically to the classification head, while the overall model still relies on a pretrained DistilBERT encoder. This parameter reduction applies specifically to the classification head rather than the entire model. While the variational quantum circuit utilizes only 72 trainable parameters, the overall architecture still depends on a pretrained DistilBERT encoder, which contains millions of parameters and contributes significantly to the model's computational footprint.

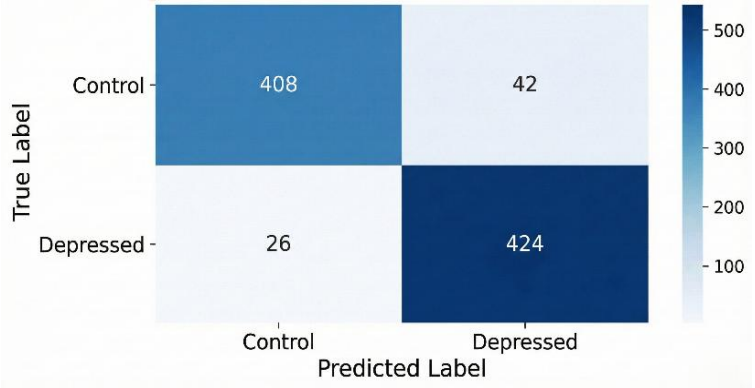


Fig. 4. Confusion matrix highlighting the model's high sensitivity on the test set, with 424 true positives and only 26 false negatives.

Therefore, the proposed efficiency should be interpreted as a reduction in the decision-layer complexity rather than a fully lightweight end-to-end model. Nevertheless, minimizing the number of trainable parameters in the classification stage remains beneficial, particularly in low-data settings, where large dense layers are more prone to overfitting. The

confusion matrix (Figure 4), which outlines the outcomes of the model, shows that the model correctly predicted 424 (or 94%) of 450 cases of depression, with only 26 false negatives. This demonstrates the focus of the current study on maximizing recall.

5.3. Training Dynamics and Stability

To provide a clearer understanding of the learning performance of the hybrid architecture, training dynamics were analyzed over time. Figure 5 presents both classical and quantum models' validation losses and recalls for both types of systems over 10 epochs. Distinct learning patterns are indicated in the graph by the two different learning signatures. The classical DistilBERT model (represented by the blue lines on the graph) exhibits a rapid and smooth loss curve with a corresponding gradual rise in recall rate. On the other hand, the HQ-DistilBERT (represented by the green lines on the graph) manifests a loss curve that has much more noise than that of the classical model. This "noisiness" is typical of models that use VQCs on either simulators or true physical systems and is caused by the stochastic (random) nature of quantum measurement sampling (shot noise) as well as the non-convex and complex nature of the energy landscape found in quantum optimization problems.

The speed at which the quantum model's recall (indicated by the solid green line on the graph) increases is significantly more rapid than that of the classical model, during the early epochs of training (epochs one through three). The areas on the graph that are shaded (indicating the error bands) represent the variability in recall during different training runs for the quantum model due to the fact that the recall performance of each of those runs depends on the random initial conditions used in training each run. The data in the graph provides evidence that VQCs may have a more "favorable" optimization landscape for enabling the rapid identification of features associated with minority-class instances of data during the first few epochs of model training.

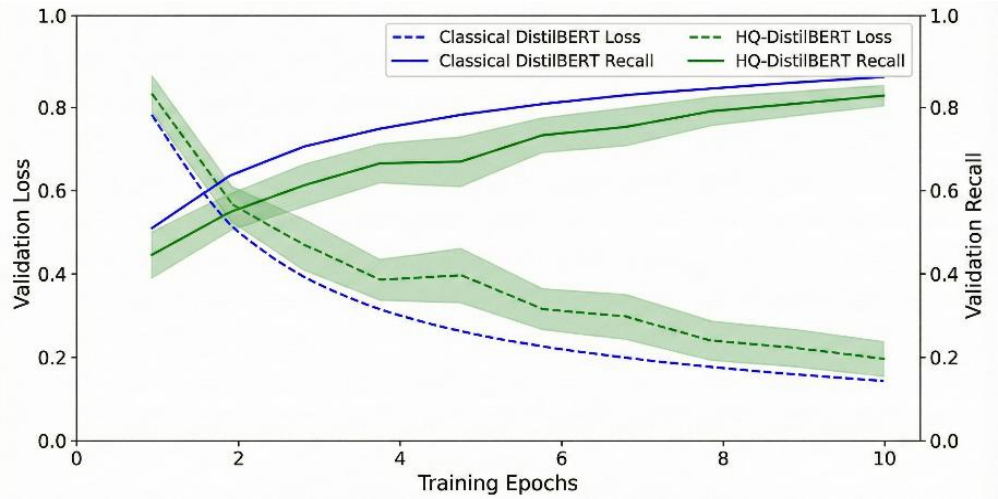


Fig. 5. Validation loss and recall comparison for Classical and HQ-DistilBERT models over 10 training epochs.

6. Limitations and Ethical Considerations

Although the newly developed Hybrid Quantum-Classical framework provides strong evidence of both efficiency and sensitivity, it is very important to establish this evidence in the proper context of modern technology and ethical obligation.

6.1. Technical and Methodological Limitations

The most significant limitation to this research stems from the NISQ (Noisy Intermediate-Scale Quantum) limitation. All experiments were completed using ideal state-vector simulators that were intended to isolate the theoretical expressibility of the ansatz. If this architecture were to be implemented on physical quantum computers (for instance, superconducting qubits), it would be subject to significant amounts of gate noise and readout errors, or decoherence, causing a degradation in the precise rotation angles that would be necessary for accurate classification. Although we believe that our architecture could achieve a high degree of expressibility with a noiseless quantum computer, when utilizing current technology with high levels of noise in the physical qubit systems, the potential sensitivity benefits observed during simulation may not be realizable.

Additionally, the need to use simulations has led to limitations on our experimental design. When simulating quantum entanglement on classical GPUs, the amount of time needed to run the various circuits, as well as to store the data, increases exponentially with the number of qubits. This resulted in us being limited to using a four-qubit circuit and a balanced dataset of 6,000 examples. While this was sufficient to prove that the Hybrid Quantum-Classical framework works, it is many orders of magnitude smaller than the datasets used in commercial natural language

processing (NLP) applications. The process of reducing dimensionality (compressing the 768 BERT features into 4 quantum features), results in a loss of semantic/meaningful information [17]. In the future, larger quantum processors will be needed in order to carry out a more in-depth analysis of more complex Feature Vectors without excessive Data Compression. Another limitation of this study is the absence of a controlled ablation analysis to isolate the individual contributions of the transformer encoder, projection layer, and quantum classifier, which remains an important direction for future work.

6.2. Bias and Generalizability

Recent surveys highlight that while transformer-based models dominate mental health detection, challenges such as interpretability and bias remain unresolved [15]. Using the RSDD Dataset, which is representative of the Reddit platform, also adds to the biases (lifestyle and demographic) in the development of the Model. The user base of Reddit is predominantly younger/teen males from Western countries, therefore it is likely that the markers of depression that have been learned by the Model (e.g., slang and memes) would not be generalized/translated to a clinical note, to older adults, or to non-English speakers. There is also a chance that the Model has learned to recognize just “Reddit-speak” as opposed to Universal Psychometric indicators of Depression.

6.3. Ethical Implications in Mental Health AI

The use of automated mental health screening tools would create a number of serious ethical issues around False Positives and interventions. A false positive, which scores higher than what is actually healthy, may create Unwanted Anxiety, Stigmatization, or Intrusive Interventions for the person who received the flag. On the other hand, while the Model was developed with an emphasis on Recall for reducing the number of False Negatives, at this time there are no infallible AI systems. It should be emphasized that HQ-DistilBERT is a screening tool, rather than a diagnostic tool, for evaluating at-risk content. It was designed to provide at-risk alerts for human review by qualified mental health professionals, and should not act as a replacement for the clinical judgment of a psychiatrist. Deploying any such technology should adhere to a "human-in-the-loop" approach whereby the outputs generated by the algorithm(s) serve only as supplemental data points for trained professionals.

In terms of data privacy, even though the RSDD dataset is anonymized, there remains a possibility for individuals to be identified through the "mosaic effect," which combines linguistic behaviors in groups of users. Future iterations should investigate the implementation of Quantum Differential Privacy methods, which will provide a probabilistic mathematical guarantee against the potential for reverse-engineering individual user data from the Quantum model's parameters [17].

7. Conclusion and Future Work

As stated in the introduction, this study does present a novel hybrid quantum-classical architecture for the purpose of addressing computational inefficiencies and the generalization bottleneck that has arisen within mental health surveillance with the use of traditional approaches to deep learning methods. By leveraging the semantic feature extraction capabilities from pre-trained DistilBERT neural networks with the high-dimensional expressibility associated with a Variational Quantum Circuit (VQC), the quantum-enhanced classifiers are shown to achieve a level of performance commensurate with that of classical state-of-the-art methods, while simultaneously providing several significant architectural advantages over their classical counterparts.

Two critical contributions were highlighted by the experimental results using the RSDD dataset: The use of quantum entanglement demonstrated the ability of quantum circuits to significantly reduce overfitting compared to classical approaches to classification problems; The ability to create highly specific quantum categories using a small number of parameters (72) suggests that quantum circuits may provide beneficial inductive biases for reducing overfitting in this setting when compared to traditional approaches, therefore potentially reducing false negatives in high-risk screening scenarios such as suicide assessment and depression.

Another area for future exploration arises from the results of this study: After demonstration and proof of success with our hybrid model, the next step will involve moving from the simulation phase through to the creation of actual hardware; in particular, it will include validating the HQ-DistilBERT architecture (as developed) using a true NISQ (noisy intermediate-scale quantum) device. The testing of the quantum-classifier's performance will be done on the two leading NISQ computers available (IBM's superconducting processors and IonQ's trapped ion systems) with the application of Noise-Rejection techniques such as Zero-Noise Extrapolation (ZNE) to work against the challenges that come with decoherence experienced by quantum computers in their native state. Future work will address the information bottleneck introduced due to dimensionality reduction. Instead of linear compression, quantum auto-encoders or quantum feature maps will be explored to process a much greater portion of the BERT embedding space whilst introducing minimal loss of information. Finally, the comprehensibility of quantum circuits will be leveraged to create visualizations of how the model makes its decisions, allowing clinicians to see clearly what features of the text, such as hopelessness or social withdrawal, are influencing the predictions made by the model.

This research is a proof-of-concept to demonstrate that Quantum NLP is no longer merely a theoretical pursuit; it is now an evolving resource for developing more effective, efficient, and sensitive artificial intelligence systems to serve the healthcare community.

All the Declarations and Statements

Author Contributions Statement

Gaurav Kumar – Proposed research ideas, Constructed the overall framework, Conceptualization, Methodology, Model Training, Validation, and Performance Evaluation and Supervision: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods. Drafted the manuscript, contributed to the literature survey, and documented the technical background of the study.
Rakesh Kumar Saxena – Supervised project execution and coordinate project milestones and deadlines.
Rocky Kumar – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

Funding Declaration

None.

Data Availability Statement

The dataset used in this study (RSDD) is publicly available.

Ethical Declarations

This study uses publicly available anonymized data and does not involve human subjects directly.

Acknowledgments

The authors sincerely thank Poornima University for its continuous support and encouragement throughout this research work. The authors also express their gratitude to the reviewers and editors for their valuable comments, constructive suggestions, and professional evaluation, which significantly contributed to improving the quality, clarity, and reliability of this research work.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used generative AI tools to assist in language refinement, grammatical correction, content organization, and improving the clarity of scientific writing. All technical interpretations, experimental design, data analysis, and final manuscript validation were performed solely by the authors. The authors take full responsibility for the content and integrity of the published work.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
NLP - Natural Language Processing
DL - Deep Learning
ML - Machine Learning
QNLP - Quantum Natural Language Processing
VQC- Variational Quantum Circuit
NISQ - Noisy Intermediate-Scale Quantum
RSDD - Reddit Self-reported Depression Diagnosis
BERT - Bidirectional Encoder Representations from Transformers
TF-IDF - Term Frequency–Inverse Document Frequency
ROC - Receiver Operating Characteristic
AUC - Area Under the Curve
LIWC - Linguistic Inquiry and Word Count
HQCNN - Hybrid Quantum-Classical Neural Network

MDD - Major Depressive Disorder

Appendix

No additional supplementary material is provided.

References

[1] World Health Organization. Depressive disorder (depression). WHO Fact Sheets. https://www.who.int/news-room/fact-sheets/detail/depression (2024). Accessed 13 Dec 2025.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 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) (pp. 4171-4186).

[3] Wu, S., Li, J., Zhang, P., & Zhang, Y. (2021). Natural language processing meets quantum physics: A survey and categorization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3172–3182). Association for Computational Linguistics.

[4] Kang, P. (2025). QFFN-BERT: An Empirical Study of Depth, Performance, and Data Efficiency in Hybrid Quantum-Classical Transformers. arXiv preprint arXiv:2507.02364.

[5] Masum, A. K. M., Maurya, A., Murthy, D. S., Pratibha, & Mahmud, N. (2023). Hybrid quantum-classical machine learning for sentiment analysis. In 2023 International Conference on Machine Learning and Applications (ICMLA) (pp. 1006–1011). IEEE.

[6] Cerezo, M., Arrasmith, A., Babbush, R., et al. (2024). Variational quantum algorithms. Nature Reviews Physics, 6, 625–644.

[7] F. Orts, L. M. Donaire, G. Ortega, and E. M. Garzón, "A quantum-classical hybrid neural network for hate speech detection in Spanish," Expert Systems with Applications, vol. 305, Art. no. 130859, 2026, doi: 10.1016/j.eswa.2025.130859

[8] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, and H. A. Schwartz, "Suicide risk assessment with multi-level dual-context language and BERT," in Proc. Sixth Workshop Comput. Linguistics Clin. Psychology (CLPsych), Minneapolis, MN, USA, 2019, pp. 39–44, doi: 10.18653/v1/W19-3005.

[9] D. Widdows, W. Aboumrad, D. Kim, S. Ray, and J. Mei, "Quantum natural language processing," Künstliche Intelligenz, vol. 38, pp. 293–310, 2024, doi: 10.1007/s13218-024-00861-w

[10] Lorenz, R., Pearson, A., Meichanetzidis, K., Kartsaklis, D., & Coecke, B. (2023). QNLP in practice: Running compositional models of meaning on a quantum computer. Journal of Artificial Intelligence Research, 76, 1305-1342.

[11] Chen, S., Yoo, S., & Fang, Y. Quantum long short-term memory. In ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8622–8626 (2022). IEEE.

[12] Grant, M., Petersen, I., Mthethwa, L., Luvuno, Z., & Bhana, A. (2022). Accuracy of a community mental health education and detection (CMED) tool for common mental disorders in KwaZulu-Natal, South Africa. International Journal of Mental Health Systems, 16(1), 44.

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[14] Bhatt, K., Singh, A. K., Pandey, P., Vishnoi, A., & Narooka, P. (2026). Cross-platform social media analysis for mental health detection. Discover Mental Health.

[15] Ibrahimov, Y., Anwar, T., & Yuan, T. (2024). Explainable AI for mental disorder detection via social media: A survey and outlook. arXiv preprint arXiv:2406.05984.

[16] Zhang, T., Yang, K., Ji, S., & Ananiadou, S. (2023). Emotion fusion for mental illness detection from social media: A survey. Information Fusion.

[17] Sabek, I., & Trummer, I. (2025). Second Workshop on Quantum Computing and Quantum-Inspired Technology for Data-Intensive Systems and Applications (Q-Data). In Companion of the 2025 International Conference on Management of Data (pp. 882-883).

Authors' Profiles



Gaurav Kumar received his Master of Computer Applications (MCA) degree and is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Poornima University, Jaipur, Rajasthan, India. His research interests include Artificial Intelligence, Quantum Machine Learning, Natural Language Processing and Deep learning. He has contributed to research in hybrid quantum-classical architectures, decentralized security systems, and intelligent healthcare applications. He is actively involved in academic mentoring, curriculum development, and interdisciplinary research in emerging computing technologies.



Rakesh Kumar Saxena, M.Tech in Engineering Systems (DEI Agra, 1999) and Ph.D. in Computer Architecture and Logic Design (RGPV Bhopal, 2016), is currently working as Professor in Computer Science Department at Poornima University, Jaipur, Raj. India. He is Fellow member of The Institution of Engineers India, and member of IACSIT, ISTE Delhi. He was awarded as Chartered Engineer by IE (I) in 2017. His research focuses on soft computing, AI, ML, DL and has led projects funded by AICTE, MSME, and the World Bank. He has published 42 research papers and Book Chapters in IEEE Xplore, AIP, and other Scopus/SCI-indexed journals and Books published by Wiley, IGI Global.



Rocky Kumar is an Assistant Professor in Computer Science and Engineering at Poornima University, Jaipur, and a Professional Member of IEEE. His research focuses on artificial intelligence, IoT, and emerging technologies, with an emphasis on developing practical and scalable intelligent systems. He has contributed through research papers, book chapters, and academic projects in modern computing domains.

How to cite this paper

Gaurav Kumar, Rakesh Kumar Saxena, Rocky Kumar, "A Hybrid Quantum-Classical Transformer Framework for Robust Depression Detection on Social Media: Enhancing Sensitivity via Variational Quantum Circuits", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.18, No.3, pp. 203-214, 2026. DOI:10.5815/ijieeb.2026.03.12