Towards Quantum-Based Detection of Adversarial Attacks in Natural Language Processing Systems

Beatrice Casey, Kaia Damian, Andrew Cotaj, Joanna C. S. Santos

University of Notre Dame
Notre Dame, IN, USA.
Email: {bcasey6, kdamian, acotaj, joannacss}@nd.edu

***Abstract*—Adversarial attacks pose a growing threat to Natural Language Processing (NLP) models by stealthily altering inputs to induce misclassifications or incorrect outputs from models. Classical detection techniques can struggle to identify them reliably due to high computational complexity, limited generalization, and susceptibility to adaptive attacks. To address these issues, we examine how quantum computing can be used to enhance the feature representations of given samples. We present an approach for detecting adversarial inputs to NLP models using Projected Quantum Kernels (PQK). Our approach transforms classical text embeddings into quantum states, evolves them to extract feature relationships, and projects them back to the classical space, where a Support Vector Machine (SVM) provides the final classification. On a dataset of 10,896 samples from the AG News dataset, this approach achieves an F1-score of 82% in its best configuration, demonstrating that quantum features are a viable method for adversarial detection. Our experiments also show that increasing quantum circuit complexity consistently degrades performance, indicating that careful, restrained circuit design is important to realize the benefits of quantum-enhanced detection.**

***Index Terms*—Quantum Computing, Natural Language Processing, Adversarial Attacks**

I. Introduction

Deep learning models for natural language processing (NLP) have become integral to systems ranging from search engines to virtual assistants. However, *adversarial attacks* threaten their security by imperceptibly altering inputs to fool models [1]. For example, an attacker could manipulate characters in a harmful post so that a model misclassifies it as safe content. Such attacks are especially stealthy due to the complex nature of human language, which is filled with idioms, metaphors, and context-dependent meanings [2]. Prior works [3], [4] have found that classical techniques struggle to detect adversarial attacks due to high computational complexity, limited generalization, and susceptibility to adaptive attacks.

These limitations motivate exploring different representational tools. Quantum feature maps embed classical data into an exponentially large Hilbert space, where entanglement between qubits can encode correlations among input features that are difficult to compute classically [5], [6]. This richer representation offers a promising basis for separating benign and adversarial samples, whose differences are often too subtle to distinguish in the original feature space.

In this poster, we present an approach which uses quantum features to train a classical Support Vector Machine (SVM). Quantum feature maps can expose hidden feature relationships that better separate benign from adversarial samples. We use Projected Quantum Kernels (PQK) to generate quantum-enriched vectors that are usable by classical classifiers.

II. Our Approach

As shown in Figure 1, the two key components of our approach are Projected Quantum Kernels (PQK) [5], [6] and Support Vector Machines (SVM) [7].

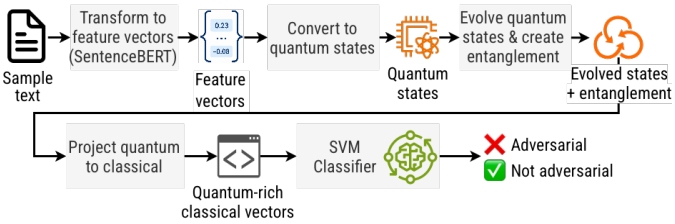

Fig. 1. Overview of our approach

Given a sample input text, we first generate classical embeddings using SentenceBERT [8]. We choose SentenceBERT because it produces lower-dimensional embeddings (384 dimensions) compared to alternatives such as BERT [9] (768 dimensions). Dimensionality is a key consideration in our approach: embeddings must be reduced to the number of available qubits, and given current hardware limitations, this reduction is severe. Thus, starting with fewer dimensions preserves more of the original information.

We then convert these features to quantum states. We apply Principal Component Analysis (PCA) to reduce the feature dimension to the target number of qubits, map them onto `ZZFeatureMaps` [10], and entangle the states using a set number of *repetitions*. We then measure the quantum states, collapsing and *projecting* them back to the classical space, leaving feature vectors that are usable by an SVM classifier, which classifies the sample as *adversarial* or *benign*.

We chose an SVM for two reasons. First, it is a long-standing method for classification problems [11]. Second, PQK returns a kernel matrix, and SVMs operate directly on kernel matrices, making the two approaches naturally compatible.

TABLE I
OUR FRAMEWORK'S DETECTION PERFORMANCE USING DIFFERENT CONFIGURATIONS. REPORTED PRECISION/RECALL/F1-SCORES ARE FOR THE POSITIVE (ADVERSARIAL) CLASS.

Reps	Qubits	Entanglement	Acc.	Prec.	Rec.	F1
1	10	Pairwise	0.82	0.85	0.77	0.81
2	**10**	**Pairwise**	**0.83**	**0.86**	**0.78**	**0.82**
2	10	Circular	0.59	0.59	0.56	0.58
2	10	Full	0.51	0.51	0.48	0.50
2	15	Pairwise	0.81	0.83	0.78	0.81
3	10	Pairwise	0.80	0.83	0.75	0.79
5	10	Pairwise	0.77	0.84	0.67	0.74
24	10	Pairwise	0.64	0.66	0.57	0.61

III. RESULTS

We evaluate our framework on IBM's simulator using a subset of the AG News dataset [12]. We generate adversarial examples with the TextAttack framework [13] using TextFooler [14] as the attack method, which perturbs characters within the input string. TextAttack runs the attack against a target model and verifies whether the model was successfully fooled. We retain the adversarial examples that successfully fooled the model alongside the original version of the text, labeled accordingly as benign or adversarial. In our experiments, we tested various combinations to find the best configuration for this task. Table I shows our results from this testing. Our best results come from using 2 repetitions, 10 qubits, and pairwise entanglement, with **83%** accuracy, **86%** precision, **78%** recall, and **82%** F1-score.

Two trends emerge from these results. First, circuit depth and detection quality are inversely related beyond 2 repetitions: the F1-score degrades from 82% at 2 repetitions to 61% at 24 repetitions, consistent with exponential concentration in quantum kernels [6], [15] in which deeper and more entangling feature maps cause kernel values between distinct inputs to converge toward a fixed value, collapsing on the geometric separation that downstream classifiers rely on. Second, entanglement density has an even stronger effect: holding depth fixed, moving from pairwise to circular to full entanglement drops the F1-score from 82% to 50%, indicating that over-entangled feature maps wash out the class-discriminative structure that sparser maps preserve. Increasing the number of qubits from 10 to 15 had negligible impact, suggesting that depth and entanglement are the dominant levers for this task.

IV. CONCLUSION

We presented a framework that uses Projected Quantum Kernels (PQK) to detect adversarial inputs to NLP models. Our approach demonstrates that quantum features can effectively distinguish benign from adversarial samples, achieving an F1-score of 82% on AG News against TextFooler attacks. Our results also surface a key empirical finding: greater quantum circuit complexity, in both depth and entanglement density, degrades detection performance rather than improving it. Future work includes (1) comparing to classical baselines, (2) evaluating alternative feature maps and measurement observables, (3) testing against a broader range of adversarial attacks, and (4) executing this approach on quantum hardware.

ACKNOWLEDGMENTS

This work is partially supported by the Center for Quantum Technologies (CQT) under the Industry-University Cooperative Research Center Program at the US National Science Foundation under Grant No. 2224985.

REFERENCES

[1] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[2] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[3] E. Abdukhamidov, T. Abuhmed, J. C. Santos, and M. Abuhamad, "Advchar: Attacking interpretable nlp systems," *IEEE Transactions on Information Forensics and Security*, 2025.

[4] F. Tramer, "Detecting adversarial examples is (nearly) as hard as classifying them," in *International conference on machine learning*. PMLR, 2022, pp. 21 692–21 702.

[5] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, no. 7747, p. 209–212, Mar. 2019. [Online]. Available: http://dx.doi.org/10.1038/s41586-019-0980-2

[6] 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 Communications*, vol. 12, no. 1, p. 2631, 2021.

[7] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," *Advances in neural information processing systems*, vol. 9, 1996.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[10] "ZZFeatureMap (latest version) — IBM Quantum Documentation — quantum.cloud.ibm.com," https://quantum.cloud.ibm.com/docs/en/api/qiskit/qiskit.circuit.library.ZZFeatureMap, [Accessed 09-10-2025].

[11] B. Casey, J. C. Santos, and G. Perry, "A survey of source code representations for machine learning-based cybersecurity tasks," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–41, 2025.

[12] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[13] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 119–126.

[14] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," 2020. [Online]. Available: https://arxiv.org/abs/1907.11932

[15] S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, "Exponential concentration in quantum kernel methods," *Nature Communications*, vol. 15, no. 1, p. 5200, 2024.