IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

Tamil Named Entity Recognition Using Classical Machine Learning, Quantum Feature Processing, and a Simplified QNLP-Inspired Component

Fexi Maxwell G[1], Paavai Anand G[2], Sridevi S[3]

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India[1-3]

Abstract: Tamil Named Entity Recognition (NER) is an important natural language processing task for identifying persons, organizations, and locations in Tamil text. This paper presents a preliminary integrated workflow combining classical machine learning (ML), quantum machine learning (QML)-inspired feature processing, and a simplified quantum natural language processing (QNLP)-inspired component. Tamil text was processed using XLM-RoBERTa, and the resulting representations were reduced using Principal Component Analysis. Classical features, simulated quantum features, and a simplified QNLP-inspired representation were evaluated with logistic-regression classifiers. In the prototype experiment, the classical baseline achieved the highest accuracy and weighted F1-score. The results demonstrate the feasibility of constructing a hybrid workflow, but they do not establish quantum advantage. The quantum circuit used fixed random parameters, and the QNLP component was a simplified proof-of-concept representation.

Keywords: Tamil NLP, Named Entity Recognition, Machine Learning, Quantum Machine Learning, Quantum Natural Language Processing, XLM-RoBERTa

I. INTRODUCTION

Named Entity Recognition identifies and classifies entities such as persons, organizations, and locations. It is useful for information extraction, search, question answering, and knowledge-base construction. Tamil presents challenges because of its rich morphology, variation in word forms, flexible syntax, and comparatively limited annotated NLP resources. These challenges motivate the investigation of both strong contextual language representations and alternative computational approaches.

Recent transformer models provide contextual representations that can support multilingual NLP tasks. At the same time, QML investigates the use of quantum circuits in machine-learning workflows, while QNLP explores the representation of linguistic structures using quantum systems. This study combines these ideas in a preliminary experimental pipeline and compares it against a classical baseline.

II. LITERATURE SURVEY

Early multilingual Named Entity Recognition research demonstrated the feasibility of cross-lingual name tagging for a large number of languages. Pan et al. [1] investigated cross-lingual name tagging and linking across 282 languages, providing an important foundation for multilingual NER and resource development for languages with limited annotated data.

Transformer-based multilingual representations have further improved the ability to process text across languages. Conneau et al. [2] introduced XLM-RoBERTa, a cross-lingual representation learning approach that provides contextual representations for multilingual NLP tasks. Such representations are particularly relevant to Tamil NER because contextual information can help address variation in word forms and linguistic structure.

Quantum Natural Language Processing (QNLP) provides an alternative perspective for representing linguistic meaning and compositional structure. Coecke et al. [3] presented mathematical foundations for a compositional distributional model of meaning, while Zeng and Coecke [4] investigated quantum algorithms for compositional natural language processing. These works establish the conceptual basis for representing relationships between linguistic components using quantum-inspired or quantum computational models.

Lorenz et al. [5] demonstrated practical approaches for running compositional QNLP models on quantum computers, showing the connection between theoretical QNLP models and quantum implementations. In addition, Qiskit Machine

 This work is licensed under a Creative Commons Attribution 4.0 International License

Learning provides tools for experimenting with quantum machine-learning algorithms [6], while scikit-learn provides established classical machine-learning methods that can be used as baseline classifiers [7].

Although these studies provide strong foundations for multilingual NER, multilingual representation learning, QML, and QNLP, their direct combination for Tamil NER remains comparatively unexplored in the scope of this work. Therefore, the present study investigates a hybrid workflow that uses XLM-RoBERTa representations, classical machine-learning classification, simulated quantum feature processing, and a simplified QNLP-inspired component. The objective is to evaluate the feasibility of connecting these approaches rather than to claim quantum advantage.

III. METHODOLOGY

The WikiAnn Tamil dataset was used as the source of annotated tokens and NER labels. A subset of 1,000 training sentences was selected for the prototype experiment. The labels represented the outside class and beginning/inside entity tags for persons, organizations, and locations.

Each sentence was tokenized using the XLM-RoBERTa tokenizer with word-level alignment. Hidden representations were extracted from the XLM-RoBERTa model. The resulting token representations were reduced to four dimensions using Principal Component Analysis (PCA). These reduced features formed the classical feature representation.

For the quantum-feature model, the four-dimensional PCA features were mapped to rotation angles in a four-qubit simulated circuit. Hadamard gates, single-qubit rotations, and a chain of controlled-NOT (CNOT) gates were used. Measurement probabilities were collected from the simulator and used as four-dimensional quantum features. The quantum circuit parameters were fixed random values and were not optimized during training.

A simplified QNLP-inspired circuit was also constructed using three sentence-level role features representing a subject, location, and action. These values were encoded into a three-qubit circuit with rotation gates and entangling operations. Because this was a local proof-of-concept implementation, the resulting feature vector was repeated across the selected token examples. It should therefore not be interpreted as a complete token-level QNLP model.

Logistic regression was used as the final classical classifier for the classical, quantum-feature, and combined representations. The data were split into training and test portions using an 80:20 split and a fixed random seed. Accuracy and weighted F1-score were used for comparison.

IV. RESULTS AND DISCUSSION

Model	Accuracy	Weighted F1-score
Classical ML baseline	0.5275	0.3748
Quantum-feature ML	0.5250	0.3615
Combined hybrid ML-QML-QNLP	0.5225	0.3695

TABLE I. PERFORMANCE COMPARISON OF THE THREE PROTOTYPE MODELS.

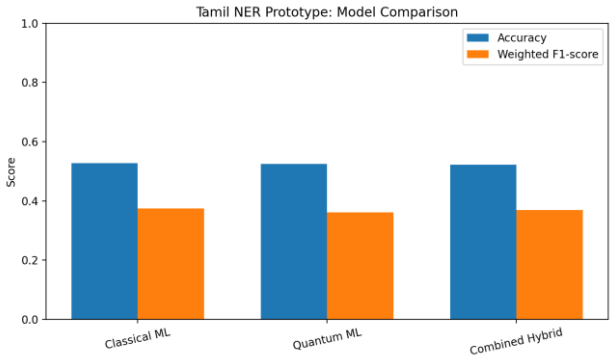


Fig. 1. Accuracy and weighted F1-score for the Tamil NER prototype.

The classical ML baseline achieved the strongest performance, with an accuracy of 52.75% and a weighted F1-score of 0.3748. The quantum-feature model achieved 52.50% accuracy and a weighted F1-score of 0.3615. The combined hybrid

model achieved 52.25% accuracy and a weighted F1-score of 0.3695. Thus, neither quantum-enhanced approach outperformed the classical baseline in this experiment.

The small differences should not be interpreted as evidence that quantum methods are ineffective in general. Instead, they reflect the limited scope of this prototype, including the use of fixed random quantum parameters, a small data subset, reduced features, and a simplified QNLP-inspired representation.

V. LIMITATIONS

The experiment used only a subset of the available data. The PCA step reduced the transformer representations to four dimensions, potentially discarding useful information. The quantum circuit parameters were not trained, so the quantum model was a feature-transformation demonstration rather than a fully optimized variational quantum model.
The QNLP component was sentence-level and simplified, with the same feature vector repeated across tokens; therefore, it was not a complete token-level compositional QNLP system. Finally, the evaluation focused on token-level accuracy and weighted F1-score rather than entity-level NER metrics.

VI. CONCLUSION AND FUTURE WORK

This paper presented a preliminary ML-QML-QNLP-inspired workflow for Tamil NER. The classical baseline performed slightly better than the quantum-feature and combined hybrid models. The work nevertheless demonstrates how transformer representations, simulated quantum circuits, and classical classifiers can be connected in a single experimental pipeline.

Future work should use the complete dataset, stronger NER architectures, trained variational quantum parameters, token-specific linguistic representations, genuine compositional QNLP models, entity-level precision/recall/F1 evaluation, and experiments on real quantum hardware. These improvements are necessary before making claims about quantum advantage or practical superiority.

VII. REPRODUCIBILITY NOTE

The experiments were developed in Google Colab using the WikiAnn Tamil dataset, XLM-RoBERTa representations, scikit-learn classifiers, Qiskit simulation, and Python-based evaluation. The numerical results reported in Table I are the values obtained from the completed prototype run.

REFERENCES

[1]. X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, “Cross-lingual name tagging and linking for 282 languages,” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1946–1958, 2017, doi: 10.18653/v1/P17-1178.

[2]. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, W. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.

[3]. B. Coecke, M. Sadrzadeh, and S. Clark, “Mathematical foundations for a compositional distributional model of meaning,” Linguistic Analysis, vol. 36, no. 1–4, pp. 345–384, 2010.

[4]. W. Zeng and B. Coecke, “Quantum algorithms for compositional natural language processing,” arXiv:1608.01406, 2016.

[5]. R. Lorenz, A. Pearson, K. Meichanetzidis, D. Kartsaklis, and B. Coecke, “QNLP in practice: Running compositional models of meaning on a quantum computer,” arXiv:2102.12846, 2021.

[6]. Qiskit Community, “Qiskit Machine Learning Documentation: Quantum Neural Networks and Quantum Machine Learning Algorithms,” 2026.

[7]. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

 This work is licensed under a Creative Commons Attribution 4.0 International License