Q2

by Rishu Kumar

Submission date: 15-Dec-2025 04:35PM (UTC+0530)
Submission ID: 2846708289
File name: q2.pdf (2M)
Word count: 6103
Character count: 38412

Q-OmniNet: Quantum-Inspired Hybrid Tensor Networks with Quantum Kernel Estimation for Polysemy-Aware Multimodal Affective Analysis of Code-Mixed Hindi–English Text

Rishu Kumar[1], Prof. Rajiv Misra[2], and Prof. T. N. Singh[3]

***Abstract*—The rapid growth of Hinglish communication on social media has introduced complex linguistic, cultural, and multimodal patterns that conventional affective computing systems struggle to handle. While deep learning and transformer-based models perform well on monolingual and grammatically consistent text, they are limited in code-mixed Hinglish settings. These models rely on deterministic, single-meaning embeddings that inadequately represent polysemy, fail to capture context-dependent polarity shifts across Hindi–English transitions, and treat multimodal signals—such as text, emojis, and visuals—as weakly coupled rather than jointly dependent. Consequently, emotional nuance, irony, and ambiguity common in real-world Hinglish interactions are often misinterpreted.**

To overcome these limitations, this work proposes Q-OmniNet, a quantum-inspired hybrid tensor network that models affective meaning as contextual superposition and captures non-separable (entangled) relationships across modalities. The framework integrates (i) a tensor-network-based contextual encoder to simulate multimodal entanglement, (ii) complex-valued embeddings to encode semantic magnitude and contextual phase, and (iii) quantum kernel estimation to model higher-order affective similarity via quantum-state overlap. Together, these components provide a richer representation of Hinglish emotion dynamics than classical or transformer-based approaches.

Experimental results on real-world multimodal Hinglish datasets show substantial improvements. Q-OmniNet achieves accuracies of 0.89 for emotion classification, 0.93 for sentiment analysis, and 0.97 for sarcasm detection, outperforming traditional baselines by 15–25% and transformer models by 8–14%. The quantum-hybrid variant demonstrates superior robustness under noisy and ambiguous conditions, achieving an average accuracy of 0.9275 compared to 0.7617 for conventional models. These findings highlight the effectiveness of quantum-inspired architectures for context-sensitive, multilingual, and multimodal affective understanding.

***Index Terms*—Quantum Computing, QTN, QKM, Transformer, LLMs, Multimodal Affective Computing.**

I. Introduction

IN the digital age, social media has become a lively place for people to share their feelings, cultures, and opinions. One of the most interesting language trends on these platforms is code-mixed communication, where users mix several languages into one message without any problems. Hindi–English code-mixing, commonly referred to as Hinglish, is one of the most common types of this phenomenon in South Asia. This hybrid linguistic form mirrors authentic bilingual speech patterns and cultural identity but presents significant obstacles for computational affective analysis owing to its informal structure, erratic grammar, and considerable contextual variability [**?**],[**?**].

Sentiment analysis, emotion recognition, and sarcasm detection are all examples of traditional affective computing frameworks that were made for text that was only in one language and didn't have many grammatical rules. When used on code-mixed data, their performance drops a lot because linguistic cues in one language often change or flip the affective polarity in another [**?**]. Also, Hinglish communication is naturally multimodal, often using emojis, GIFs, or pictures to add to the emotional context. The interplay of these modalities generates intricate emotional signals that cannot be accurately represented through unimodal or single embedding frameworks.

Deep learning and transformer-based architectures have achieved remarkable success in natural language understanding; however, they remain fundamentally limited when addressing polysemy, contextual ambiguity, and cross-modal dependencies in informal communication. Standard transformer embeddings encode each token as a single deterministic vector, which does not reflect the superposed nature of meaning where a word or phrase can simultaneously convey multiple emotional states depending on context [**?**],[**?**]. Likewise, traditional multimodal fusion techniques frequently depend on feature concatenation or attention-based weighting, which fail to adequately capture non-separable dependencies between modalities like text and emoji. Consequently, nuanced affective subtleties such as irony, humor [**?**], and ambivalence are frequently misclassified or inadequately represented.

To tackle these issues, this study presents Q-OmniNet, a quantum-inspired hybrid tensor network framework aimed at polysemy-aware multimodal affective analysis of code-mixed Hindi–English text. The fundamental assertion of Q-OmniNet is that emotional significance in natural language communication displays characteristics similar to those found in quantum systems namely, superposition (the coexistence

[1]Rishu Kumar, Department of Computer Science & Engineering, Indian Institute of Technology Patna, India. Email: rishu_2422cs04@iitp.ac.in
[2]Prof. Rajiv Misra, Department of Computer Science & Engineering, Indian Institute of Technology Patna, India. Email:rajivm@iitp.ac.in
[3]Prof. T. N. Singh, Department of Civil and Environmental Engineering, Indian Institute of Technology Patna, India. Email:tnsiitb@gmail.com

of multiple meanings) and entanglement (the interdependence among modalities)[?],[?],[?]. By applying these principles in a traditional computational environment, Q-OmniNet presents an innovative representational framework that combines quantum-inspired embeddings, tensor-based contextual encoding, and quantum kernel estimation for multitask affective inference[?],[?].

In the proposed framework, linguistic, paralinguistic (emoji), and visual modalities are initially synchronized into a common latent space via multimodal encoders. In a Hilbert space, each token or feature is then represented as a complex-valued amplitude vector. The real and imaginary parts of the vector together encode semantic and contextual phases [?], [?],[?]. The resulting superimposed representations allow the model to capture simultaneous emotional interpretations. A Quantum Tensor Network (QTN) module models contextual and cross-modal dependencies through tensor contraction operations that mimic entanglement[?]. Finally, a Quantum Kernel Estimation (QKE) layer uses quantum-state overlap to find similarities between multimodal emotional states. This lets you tell the difference between different emotions in tasks [?],[?] like recognizing emotions, analyzing sentiment, and finding sarcasm.

This work enhances the domains of affective computing and multimodal comprehension in several significant aspects:

1. Quantum-Inspired Affective Modeling: Proposes an innovative theoretical framework utilizing quantum concepts—superposition and entanglement—as computational analogies to represent polysemy and multimodal emotional interconnection in code-mixed social media interactions[?],[?].

2. Hybrid Tensor Network Design: This idea combines tensor network representation leaming with quantum-inspired embeddings to create a hybrid architecture that allows for high-order contextual reasoning without making things too complicated[?].

3. Multimodal and Multitask Integration: Develops a unified model that simultaneously performs emotion, sentiment, and sarcasm analysis by exploiting shared affective structures across modalities[?].

4. Empirical Advancement: Demonstrates the effectiveness of Q-OmniNet over classical and transformer baselines through extensive experiments on real-world Hinglish multimodal datasets, showing substantial improvements in accuracy and robustness[?],[?].

By bridging concepts from quantum information theory and deep learning, Q-OmniNet establishes a new direction for multimodal affective computing—one that treats emotion understanding as a phenomenon of distributed, context-dependent meaning rather than static classification. The suggested framework not only improves the ability to recognize emotions in code-mixed digital communication, but it also helps with the bigger goal of creating quantum-inspired cognitive architectures that can model the complex, intertwined nature of human emotion and language.

A. Motivation

Even though multimodal affective computing has advanced significantly, current approaches are still unable to accurately capture the ambiguity and contextual fluidity found in actual Hinglish communication. Conventional methods rely on fixed, deterministic embeddings that are unable to capture how modalities like text, emojis, and images collectively influence meaning or preserve a variety of possible emotional interpretations. The necessity for a modeling framework that can depict emotional meaning as context-dependent and intrinsically ambiguous is highlighted by this gap. Such a foundation is made possible by quantum-inspired representations, which allow for the superposed and interdependent encoding of multimodal cues. This inspires the development of Q-OmniNet, which utilizes these principles to more properly depict the dynamic and interwoven nature of human affect.

B. Problem Statement

The increasing use of code-mixed Hinglish on social media presents emotional expressions that are highly variable, multimodal, and context dependent. These posts frequently combine Hindi and English tokens with emojis and visual cues, creating emotional signals that cannot be reliably interpreted using existing models. Current deep learning and transformer-based systems rely on fixed token embeddings and largely independent multimodal fusion mechanisms, which prevents them from capturing multiple possible meanings of a word or the joint influence of text, emoji, and image modalities. As a result, these models struggle with ambiguity, polarity shifts, and subtle expressions such as mixed sentiment or sarcasm.

This work addresses this gap by developing a quantum-inspired hybrid tensor network that represents emotional meaning as a contextual superposition and models cross-modal dependencies as entangled interactions. This framework enables a more expressive and context-aware approach to affective analysis in code-mixed digital communication.

II. Literature Review

Quantum-inspired neural models, which mimic Hilbert-space formalisms on classical hardware, have made significant strides in tandem with hybrid architectures. *Shi et al*. presented two end-to-end quantum-inspired deep neural networks that outperformed classical baselines on several text-classification benchmarks by modeling linguistic polysemy and contextual interference using complex-valued word embeddings[?]. Their later work proposed pretrained quantum-inspired embeddings (QPFE) that encode superposition-based word semantics and integrate with ERNIE, achieving superior performance in sentiment classification and word sense disambiguation compared to both classical and earlier quantum-inspired approaches[?].

Research at the interface of quantum machine learning (QML) and natural language processing (NLP) has rapidly increased in recent years, motivated by the limits of classical deep neural networks and the rising potential of quantum representations. Early breakthroughs integrated variational quantum circuits (VQCs) with multimodal and multitask learning. Phukan et al., for example, demonstrated the ability of quantum circuits to encode cross-modal dependencies with significantly fewer parameters than classical systems by proposing a

hybrid quantum–classical neural network that uses superposition, entanglement, and interference to jointly model sarcasm, sentiment, and emotion across text–audio–visual modalities [**?**].

Efforts to improve the scalability and physical realizability of QML have developed effective classical–quantum collaboration structures. *L'Abbate et al.* presented co-TenQu, integrating tensor-network–based compression with VQC-based classification, greatly decreasing qubit usage while boosting accuracy over state-of-the-art quantum baselines. Robust end-to-end training spanning classical and quantum components was further made possible by their fidelity-driven feedback loop[**?**].

Another area of research focuses on quantum models for resource-constrained NLP applications. *Yu et al.* created batch-upload quantum recurrent neural networks (BUQRNNs) to solve dimensionality limits in NISQ devices, demonstrating enhanced performance in low-resource Bengali sentiment classification while decreasing model complexity through optimized VQC design[**?**].

Memory-efficient quantum computation has also been researched. *Khan et al.* presented EP-PQM, a parametric probabilistic quantum memory that uses label encoding to decrease the amount of necessary qubits and gate operations. Their technique dramatically decreases circuit depth and qubit count compared to classical PQM, hence enabling more scalable quantum machine learning on NISQ devices[**?**].

Quantum concepts have been included into complex-valued neural architectures aside of NLP-specific applications. *Zhang et al.* developed a complex-valued neural network quantum language model (C-NNQLM) that produces differentiable density matrices from end-to-end embeddings, improving question answering, document retrieval, and text classification. Their results reveal that storing word order via complex phases gives significant advantages over real-valued quantum-inspired models and classical baselines[**?**].

Finally, reliability issues have prompted research on error propagation in quantum layers as quantum neural models proliferate. *Vallero et al.* examined logical-shift faults in quantum convolutional neural networks, demonstrating the need for reliable error-tolerant QML designs and demonstrating that even little changes to qubit states can result in significant misclassification rates[**?**].

Overall, present work together underlines both the potential and the practical constraints of quantum and quantum-inspired NLP models. They demonstrate better representational capacity and interpretability using quantum principles, although encounter limits relating to qubit availability, circuit depth, and noise in NISQ devices. These restrictions inspire the construction of more scalable, interpretable, and parameter-efficient quantum–classical structures, aligning with the objectives of the present research.

III. Theoretical Background and Preliminaries

The proposed Q-OmniNet framework builds upon mathematical concepts derived from quantum information theory and tensor-based representation learning, which are adapted here to enhance affective understanding in multimodal and code-mixed linguistic contexts [**?**],[**?**],[**?**]. This section introduces the fundamental principles and notations that underpin the proposed methodology: the Hilbert-space representation of information, the quantum-inspired principles of *superposition* and *entanglement*, and the tensor network formalism used for contextual encoding.

A. Hilbert space representation

In quantum theory the state of a system is represented as a normalized complex vector in a Hilbert space $\mathcal{H} \subseteq \mathbb{C}^d$:

$$|\psi\rangle \in \mathcal{H}, \qquad \langle\psi|\psi\rangle = 1. \tag{1}$$

Each component of $|\psi\rangle$ corresponds to a basis configuration and its complex amplitude encodes magnitude and phase. The inner product between two states measures their overlap:

$$\langle\phi|\psi\rangle = \sum_{k=1}^{d} \phi_k^* \psi_k. \tag{2}$$

In Q-OmniNet, embeddings (textual, emoji, visual) are interpreted as state vectors in such a Hilbert space, where magnitude may represent intensity and phase encodes contextual variation[**?**],[**?**],[**?**],[**?**].

B. Superposition: modeling polysemy

The superposition principle permits a system to occupy multiple states simultaneously [**?**],[**?**]:

$$|\Psi\rangle = \sum_{i=1}^{N} \alpha_i |\psi_i\rangle, \qquad \sum_i |\alpha_i|^2 = 1, \tag{3}$$

where $\alpha_i \in \mathbb{C}$ are probability amplitudes. In linguistic modeling this concept is used to represent *polysemy*: a token may have multiple contextual meanings (emotional or semantic) which coexist until disambiguated by context[**?**],[**?**],[**?**]. Q-OmniNet leverages superposed representations to preserve multiple candidate interpretations of Hinglish tokens[**?**],[**?**].

C. Entanglement: modeling cross-modal dependency

Entanglement describes joint states that are not separable into independent subsystem states:

$$|\Psi_{AB}\rangle \neq |\psi_A\rangle \otimes |\psi_B\rangle. \tag{4}$$

This models non-separable correlations between modalities: for example, the emotional effect of an emoji may depend heavily on the accompanying text[**?**],[**?**]. Q-OmniNet operationalizes this notion via tensor interactions that capture high-order inter-modal dependencies[**?**],[**?**].

D. Tensor network formalism

Tensor networks (TNs) offer an efficient representation for high-dimensional, structured dependencies by arranging tensors [**?**],[**?**] and contracting them along shared indices. Given a sequence of state vectors $|\psi_1\rangle, \ldots, |\psi_N\rangle$, a tensor-network–based contextual encoding can be written as:

$$T_{\text{TN}} = \text{contract}\left(U_1|\psi_1\rangle, U_2|\psi_2\rangle, \ldots, U_N|\psi_N\rangle\right), \tag{5}$$

where U_i are learnable transformation tensors (often constrained to be orthogonal/unitary-like for stability). The contraction operation aggregates local correlations into a compact global representation that reflects entanglement-like relationships among tokens and modalities[?], [?], [?].

E. Quantum kernel similarity

To measure similarity between two encoded representations x_i and x_j in a quantum-inspired feature space, Q-OmniNet uses a quantum-like kernel defined by state overlap[?], [?]:

$$K_Q(x_i, x_j) = \left| \langle \Phi_Q(x_i) | \Phi_Q(x_j) \rangle \right|^2, \tag{6}$$

where $\Phi_Q(\cdot)$ maps inputs into the quantum feature space. In classical simulations the kernel is approximated by the squared magnitude of the complex inner product of amplitude embeddings:

$$K_Q(x_i, x_j) \approx \left| \psi_i^\dagger \psi_j \right|^2. \tag{7}$$

This quantum-inspired kernel captures nonlinear and contextual similarities that are useful for distinguishing subtle affective states such as sarcasm and ambivalence[?], [?], [?], [?].

The theoretical framework used by Q-OmniNet rests on the following elements:

- Hilbert space representation: Embeddings interpreted as normalized complex state vectors (Eq. 1–2)[?], [?].
- Superposition: Polysemy modeled as superposed amplitude states (Eq. 3)[?], [?], [?].
- Entanglement: Cross-modal dependencies captured via tensor contractions (Eq. 4, 5)[?], [?].
- Quantum kernel: Similarity measured via state overlaps for nonlinear affective discrimination (Eq. 31–7)[?], [?].

These preliminaries supply the mathematical vocabulary used in the following section, where we present the computational design and full mathematical formulation of the Q-OmniNet architecture.

To facilitate readability, Table I provides a consolidated list of important symbols and notations employed in the Q-OmniNet framework.

TABLE I
LIST OF IMPORTANT SYMBOLS AND THEIR DESCRIPTIONS USED IN THE Q-OMNINET FRAMEWORK.

Symbol	Description
$X = \{X_t, X_e, X_i\}$	Multimodal input consisting of text (X_t), emojis (X_e), and image (X_i)
w_i	i-th token in the Hinglish text sequence
$v_i \in \mathbb{R}^d$	Contextual embedding of token w_i from transformer encoder
$L_i \in \{H, E, O\}$	Language identifier for token
$V_t \in \mathbb{R}^{N \times (d+3)}$	Text embedding matrix with language tag appended
$E_j \in \mathbb{R}^d$	Embedding vector for j-th emoji
$\tilde{E}$	Attention-weighted aggregated emoji embedding
I_{raw}	Raw visual feature from Vision Transformer / CLIP
$\tilde{I} \in \mathbb{R}^d$	Projected visual embedding
X_{multi}	Unified multimodal representation vector
ψ_k	Complex quantum-inspired embedding for feature x_k
$\|\Psi\rangle$	Quantum superposition representation
α_k	Attention-derived amplitude weight
U_i	Tensor in the Quantum Tensor Network
T_{QTN}	Contracted QTN tensor
$K_Q(x_i, x_j)$	Quantum kernel function
$\Phi_Q(\cdot)$	Quantum feature map
$\psi_i^\dagger$	Hermitian transpose
Z	Representation after QKE
$\hat{y}_{\text{emo}}$	Emotion prediction
$\hat{y}_{\text{sent}}$	Sentiment prediction
$\hat{y}_{\text{sar}}$	Sarcasm prediction
$\mathcal{L}_{\text{total}}$	Overall multitask loss
$\mathcal{L}_{\text{emo}}, \mathcal{L}_{\text{sent}}, \mathcal{L}_{\text{sar}}$	Task-specific losses
$\lambda_1, \lambda_2, \lambda_3$	Multitask loss weights
$\oplus$	Concatenation operator
$\otimes$	Tensor/Kronecker product
$\langle \phi \| \psi \rangle$	Inner product
$\|\cdot\|^2$	Modulus-squared amplitude

IV. Deep Learning Foundations of Q-OmniNet

Section 3 established the quantum-theoretic conceptual backbone of Q-OmniNet, describing how superposition, entanglement, and Hilbert-space representations motivate a non-classical view of multimodal affect[?], [?], [?]. In this section, we now operationalize those principles into an implementable neural pipeline. Specifically, we describe how standard deep learning modules (transformers for text, attention pooling for emojis, and ViT/CLIP encoders for images) serve as the substrate upon which the later quantum-inspired embedding, tensor interaction, and kernel estimation mechanisms are applied[?], [?], [?], [?]. Thus, Section 4 forms the bridge between the theoretical quantum analogy (Section 3) and the full computational formulation (Section 5).

While the theoretical design of Q-OmniNet is inspired by quantum mechanics, its computational realization is grounded in modern deep learning architectures. This section introduces the neural foundations that form the backbone of the proposed framework, describing the core components used to extract, align, and fuse multimodal features before quantum-inspired processing[?], [?].

A. Transformer-Based Textual Encoding

For the textual modality, Q-OmniNet employs a multilingual transformer encoder such as XLM-R or MuRIL to capture contextual semantics across code-mixed Hindi–English text[?], [?]. Given an input sequence $X_t = [w_1, w_2, \ldots, w_N]$, each token w_i is first embedded and then contextualized using multi-head self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{8}$$

where Q, K, and V denote the query, key, and value matrices respectively, and d_k is the dimensionality of the key vectors.

Contextual representation is refined through a series of transformer blocks that include residual connections, layer normalization, and feed-forward networks:

$$H = \text{LayerNorm}\big(H + \text{FFN}(\text{MultiHead}(H))\big), \tag{9}$$

where H represents the hidden state sequence. This architecture enables the encoder to preserve both local (token-level) and global (sentence-level) semantics [?], [?], forming the primary linguistic input for quantum-inspired embedding generation.

B. Visual and Emoji Feature Encoders

For the visual modality, a pretrained Vision Transformer (ViT) or CLIP encoder is employed to extract visual embeddings[?], [?]:

$$I_{\text{raw}} = f_{\text{vis}}(X_i) \in \mathbb{R}^{h' \times w' \times c'}. \tag{10}$$

A projection layer aligns these features with the shared latent space:

$$\tilde{I} = W_i \,\text{flatten}(I_{\text{raw}}) \in \mathbb{R}^d. \tag{11}$$

For the emoji modality, each emoji e_j is represented through an affective embedding derived from sentiment lexicons and emotion intensity scores[?], [?]:

$$E_j = f_{\text{emo}}(e_j) \in \mathbb{R}^d. \tag{12}$$

An attention pooling mechanism aggregates emoji-level information:

$$\tilde{E} = \sum_{j=1}^{M} \beta_j E_j, \qquad \beta_j = \frac{\exp(\gamma_j)}{\sum_k \exp(\gamma_k)}, \tag{13}$$

where β_j denotes the normalized attention weight for the j-th emoji[?], [?].

C. Multimodal Fusion Mechanism

Before applying quantum-inspired transformations, Q-OmniNet aligns textual, emoji, and visual features into a shared latent space through a deep multimodal fusion module[?], [?]. The unified representation is constructed by concatenating the encoded features:

$$X_{\text{multi}} = [\tilde{V}_t \oplus \tilde{E} \oplus \tilde{I}], \tag{14}$$

where $\oplus$ denotes concatenation along the feature dimension.

A multimodal attention layer then refines these features to emphasize emotionally salient cross-modal dependencies:

$$\hat{X}_{\text{multi}} = \text{softmax}(QK^\top)V. \tag{15}$$

This process ensures that emotionally relevant interactions across modalities (e.g., between text and emoji) are amplified prior to quantum encoding[?], [?].

D. Integration with Quantum-Inspired Modules

The deep learning backbone thus establishes a robust semantic foundation upon which the quantum-inspired components of Q-OmniNet operate. Transformer and attention mechanisms capture hierarchical linguistic and visual features, while the subsequent Quantum Tensor Network (QTN) and Quantum Kernel Estimation (QKE) modules extend these representations to model higher-order semantic superpositions and entangled affective dependencies[?], [?], [?], [?]. In this way, Q-OmniNet combines the expressive capacity of deep neural architectures with the contextual depth of quantum-inspired representation learning, achieving an enriched modeling of emotion, sentiment, and sarcasm in code-mixed digital communication[?], [?]. As shown in Fig. 1, the modality-specific embeddings are projected into a shared latent space and fused to form X_{multi} , which is subsequently mapped to complex-valued quantum-inspired embeddings.

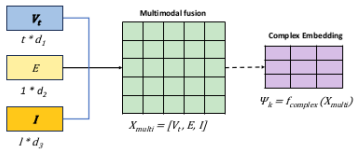

Fig. 1. Embedding pipeline used in Q-OmniNet. The text sequence, emoji token, and image patches are first encoded into matrices of dimensions $V_t \in \mathbb{R}^{t \times d_1}$, $E \in \mathbb{R}^{1 \times d_2}$, and $I \in \mathbb{R}^{\ell \times d_3}$, respectively. These modality-specific embeddings are projected into a unified feature space and concatenated to form the multimodal representation $X_{\text{multi}} = [V_t; E; I]$. The fused representation is then transformed by the complex embedding layer to produce the quantum-inspired embedding $\psi_k \in \mathbb{C}^{\ell \times d}$, which serves as the input to the subsequent QTN and QKM modules.

V. Proposed Model

This section presents the complete mathematical formulation and computational design of the proposed Q-OmniNet framework a quantum-inspired hybrid tensor network for *polysemy-aware multimodal affective analysis* of code-mixed Hindi–English (Hinglish) text. Building upon the quantum principles introduced in Section III, Q-OmniNet integrates superposition, entanglement, and tensorized contextual learning to jointly model semantic ambiguity and cross-modal emotional dependencies.

The overall computational pipeline follows five sequential stages:

1) Multimodal Input Representation and Preprocessing,
2) Quantum-Inspired Embedding Generation,
3) Quantum Tensor Network (QTN) Encoding,
4) Quantum Kernel Estimation (QKE), and
5) Multitask Affective Classification.

As illustrated in Fig. 2, Q-OmniNet integrates multimodal encoders with quantum-inspired embedding, tensor interaction, and kernel estimation modules. The overall training pipeline of Q-OmniNet is outlined in Algorithm 1.

A. Multimodal Input Representation and Preprocessing

Let a multimodal post be denoted as

$$X = \{X_t, X_e, X_i\}, \tag{16}$$

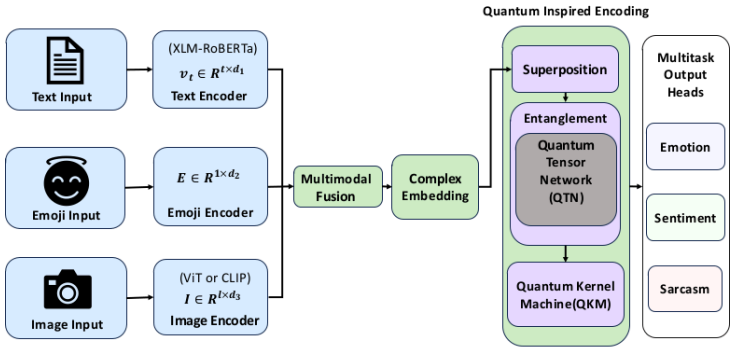

Fig. 2. Model Overview

where X_t is the text sequence, X_e the emoji sequence, and X_i the corresponding image. The model performs normalization, tokenization, and feature alignment to ensure consistent multimodal inputs.

1) Textual Representation: Given a Hinglish sentence $X_t = [w_1, w_2, \ldots, w_N]$, each token w_i is encoded using a multilingual transformer (e.g., XLM-R):

$$v_i = f_{\text{emb}}(w_i) \in \mathbb{R}^d \tag{17}$$

A 3-dimensional one-hot vector $L_i \in \{H, E, O\}$ represents the language tag (Hindi, English, Other) and is concatenated with the embedding:

$$v_i' = [v_i; \text{onehot}(L_i)] \in \mathbb{R}^{d+3} \tag{18}$$

The resulting embedding matrix is:

$$V_t = [v_1', v_2', \ldots, v_N']^\top \in \mathbb{R}^{N \times (d+3)} \tag{19}$$

2) Emoji Representation: Each emoji e_j is mapped to an affective vector using a sentiment lexicon:

$$E_j = f_{\text{emo}}(e_j) \in \mathbb{R}^d. \tag{20}$$

Attention pooling aggregates emoji information as:

$$\tilde{E} = \sum_{j=1}^{M} \beta_j E_j, \qquad \beta_j = \frac{\exp(\gamma_j)}{\sum_k \exp(\gamma_k)} \tag{21}$$

3) Visual Representation: For the image modality, a pretrained Vision Transformer or CLIP encoder extracts visual features:

$$I_{\text{raw}} = f_{\text{vis}}(X_i) \in \mathbb{R}^{h' \times w' \times c'} \tag{22}$$

A projection layer aligns them with the textual space:

$$\tilde{I} = W_i \, \text{flatten}(I_{\text{raw}}) \in \mathbb{R}^d \tag{23}$$

4) Unified Multimodal Representation: All modalities are concatenated to obtain:

$$X_{\text{multi}} = [\tilde{V}_t \oplus \tilde{E} \oplus \tilde{I}] \in \mathbb{R}^{d_{\text{multi}}} \tag{24}$$

where $\oplus$ denotes concatenation. This unified vector serves as the input to the quantum-inspired embedding generator.

B. Quantum-Inspired Embedding Generation

Each multimodal feature vector $x_k \in X_{\text{multi}}$ is transformed into a complex-valued embedding:

$$\psi_k = \psi_k^R + j\psi_k^I \in \mathbb{C}^d \tag{25}$$

where ψ_k^R and ψ_k^I represent semantic and contextual (phase) components, respectively, and

$$\|\psi_k\|_2 = 1 \tag{26}$$

A post-level representation is then modeled as a *superposed quantum state*:

$$|\Psi\rangle = \sum_{k=1}^{N'} \alpha_k |\psi_k\rangle, \qquad \sum_k |\alpha_k|^2 = 1 \tag{27}$$

where attention-derived weights α_k are defined as:

$$\alpha_k = \frac{\exp(q_k)}{\sum_m \exp(q_m)}, \qquad q_k = \text{score}(\psi_k, \text{context}) \tag{28}$$

This superposed representation captures multiple coexisting emotional interpretations, enabling polysemy-aware modeling.

In this work, each token is mapped to a complex-valued representation that preserves both semantic magnitude and contextual phase. The model's ability to represent tokens in superposed complex embedding states enables it to encode multiple potential meanings of a polysemous Hinglish term

before context-based disambiguation. This significantly contributes to improved accuracy, particularly in emotion and sarcasm tasks where contextual polarity shifts are common. These superposed embeddings allow the model to delay hard semantic assignments until global context is processed.

C. Quantum Tensor Network (QTN) Encoding

The Quantum Tensor Network (QTN) encodes contextual and inter-modal dependencies through tensor contractions:

$$T_{\text{QTN}} = \text{contract}(U_1|\psi_1\rangle, U_2|\psi_2\rangle, \ldots, U_{N'}|\psi_{N'}\rangle) \tag{29}$$

where $U_i \in \mathbb{C}^{r_{i-1} \times d \times r_i}$ are transformation tensors and r_i denote bond dimensions.

The explicit contraction can be expressed as:

$$T_{\text{QTN}} = \sum_{i_1,\ldots,i_{N'}} U_1^{(i_1)} \otimes U_2^{(i_2)} \cdots \otimes U_{N'}^{(i_{N'})} \prod_{k=1}^{N'} \psi_{i_k} \tag{30}$$

This operation models *contextual entanglement*, ensuring that affective meaning from one modality influences others.

D. Quantum Kernel Estimation (QKE)

The Quantum Kernel Estimation (QKE) layer measures similarity between contextual states based on quantum state overlap:

$$K_Q(x_i, x_j) = \left|\langle \Phi_Q(x_i) | \Phi_Q(x_j) \rangle\right|^2 \tag{31}$$

where $\Phi_Q(\cdot)$ maps inputs into a quantum feature space. In classical simulation, this overlap is approximated by:

$$K_Q(x_i, x_j) \approx |\psi_i^\dagger \psi_j|^2 \tag{32}$$

The kernel values are used to construct a multitask affective inference layer.

E. Multitask Affective Classification

Let the shared representation after QKE be $Z \in \mathbb{R}^{B \times d_k}$, where B is the batch size and d_k the kernel dimension. Predictions for emotion, sentiment, and sarcasm are computed as:

$$\hat{y}_{\text{emo}} = \text{softmax}(W_{\text{emo}} Z + b_{\text{emo}}), \tag{33}$$

$$\hat{y}_{\text{sent}} = \text{softmax}(W_{\text{sent}} Z + b_{\text{sent}}), \tag{34}$$

$$\hat{y}_{\text{sar}} = \sigma(W_{\text{sar}} Z + b_{\text{sar}}), \tag{35}$$

where $\sigma(\cdot)$ denotes the sigmoid activation.

The total multitask loss function is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{emo}} + \lambda_2 \mathcal{L}_{\text{sent}} + \lambda_3 \mathcal{L}_{\text{sar}}, \tag{36}$$

with each individual task loss defined as:

$$\mathcal{L}_{\text{task}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i). \tag{37}$$

F. End-to-End Computational Flow

The complete computational process of Q-OmniNet is summarized as:

$$X \xrightarrow{\text{Preprocessing}} X_{\text{multi}} \xrightarrow{\text{Quantum Encoding}} \{|\psi_i\rangle\} \xrightarrow{\text{Tensor Contraction}} T_{\text{QTN}} \xrightarrow{\text{Quantum Kernel}} Z \xrightarrow{\text{Multitask Classifier}} \hat{Y}. \tag{38}$$

Superposition enables semantic flexibility for code-mixed tokens, while tensor entanglement captures emotional coupling across modalities.

G. Computational Variants

Three computational variations of Q-OmniNet were built in order to assess the contribution of the quantum-inspired components:

- Classical Baseline: Uses real-valued embeddings and traditional attention-based multimodal fusion, without any quantum-inspired procedures.
- Quantum-Inspired: Incorporates complex-valued embeddings and superposition-based contextual modeling to capture polysemy and contextual ambiguity, but does not include tensor-network or quantum-kernel modules.
- Quantum–Classical Hybrid: The Quantum–Classical Hybrid Variant combines the Quantum Tensor Network (QTN) and Quantum Kernel Method (QKM/QKE). It is implemented by simulating quantum-inspired processes in a classical manner without the need of quantum frameworks. These components together model entangled multimodal dependencies and quantum-like similarity.
  The comparison between these three types, shown in the Experiments section, demonstrates the effect of complex embeddings, tensor-network contextualization, and quantum kernel similarity on overall performance.

The proposed Q-OmniNet integrates the mathematical strengths of quantum formalism and tensor networks within a unified neural framework. By representing multimodal affective information as *superposed and entangled quantum-like states*, and evaluating emotional similarity through *quantum kernel overlaps*, the model achieves enhanced contextual understanding of polysemous and multimodal Hinglish communication.

VI. Model Description

A unified multimodal processing pipeline consisting of text, emoji, and picture encoders was used to develop and train the proposed Q-OmniNet framework. Next came quantum-inspired embedding modules, a tensor-network contextualizer, and a quantum kernel inference layer. An NVIDIA A100-PCIE-40GB GPU was used for all of the trials, running at a power drain of about 245 W and continuously functioning in performance state P0 during training. This design provides adequate memory capacity and computing throughput to allow complex-valued embedding operations, tensor contractions, and large-batch quantum-kernel calculations.

Python-based deep learning frameworks with mixed-precision training enabled to maximize computation efficiency

Algorithm 1 Q-OmniNet — concise training pipeline

Require: Dataset $\mathcal{D} = \{(X_t, X_e, X_i), Y\}$, encoders $f_{\text{emb}}, f_{\text{vis}}$, emoji map f_{emo}, hyperparams $d, B, E, \eta, \{r_i\}, d_k, \lambda$, Variant
Ensure: Trained model and test metrics

Preprocess $\mathcal{D}$: tokenize/transliterate, language-tag, map emojis, normalize images
Split $\mathcal{D} \rightarrow$ train/val/test
Initialize $f_{\text{emb}}, f_{\text{vis}}, W_t, W_e, W_i$;
if Variant $\neq$ CLASSICAL **then**
    initialize $g_{\text{cplx}}, \{U_i\}$, QKE
end if
for epoch $= 1$ to E **do**
    for minibatch $\mathcal{B}$ **do**
        for each sample $(X_t, X_e, X_i) \in \mathcal{B}$ **do**
            $V_t \leftarrow f_{\text{emb}}(X_t)$; append language tags
            $\tilde{E} \leftarrow \text{attn\_pool}(f_{\text{emo}}(X_e))$
            $\tilde{I} \leftarrow W_i(\text{flatten}(f_{\text{vis}}(X_i)))$
            $X_{\text{multi}} \leftarrow [\text{pool}(V_t) \oplus \tilde{E} \oplus \tilde{I}]$
            if Variant == CLASSICAL **then**
                features $\leftarrow X_{\text{multi}}$
            else
                $\psi \leftarrow g_{\text{cplx}}(X_{\text{multi}})$ ▷ complex map & normalize
                $\alpha \leftarrow \text{attn}(\psi)$; $|\Psi\rangle \leftarrow \sum_k \alpha_k |\psi_k\rangle$
            end if
            $T_{\text{QTN}} \leftarrow \text{contract}(\{U_i, |\psi_i\rangle\})$
        end for
        $K \leftarrow \text{QKE}(\{T_{\text{QTN}}\})$ ▷ approx. $|\psi_i^\dagger \psi_j|^2$
        $Z \leftarrow \text{kernel\_features}(K)$
        $\hat{Y} \leftarrow \text{classifiers}(Z)$; compute $\mathcal{L}_{\text{total}}$
        Backpropagate and update parameters
    end for
    Validate; early stop if needed
end for
Test evaluation and variant comparison

were used to create the model. During training, the textual encoder received tokenized Hinglish sequences enriched with language tags, while emoji embeddings and visual features were aligned using projection layers before multimodal fusion. The quantum-inspired embedding block normalized features into complex amplitude vectors, which were further processed by the Quantum Tensor Network. Kernel features were created using the quantum-overlap similarity function and fed to task-specific classifiers for emotion, sentiment, and sarcasm prediction.

Training was performed with mini-batches ranging from 13 to 32 samples, depending on modality configuration, and the model was optimized using Adam with a scheduled learning rate. Early halting and validation-based checkpointing were utilized to ensure steady convergence across jobs. Smooth training dynamics and repeatable performance across several runs were made possible by the A100 GPU's effective handling of large-scale kernel matrices and tensor-network contractions.

TABLE II
COMPARISON OF Q-OMNINET WITH EXISTING APPROACHES

Model	Emotion Acc.	Sentiment Acc.	Sarcasm Acc.
Q-OmniNet (Proposed)	**0.89**	**0.93**	**0.97**
Classical Baseline (Real-valued embeddings + attention)	0.74	0.78	0.77
Transformer Baseline (XLM-R)	0.81	0.86	0.83
Hybrid Quantum-Classical (Phukan et al.)	0.84	0.88	0.85

VII. EVALUATION

In the Fig. 3. The training dynamics of Q-OmniNet demonstrate stable convergence and strong generalization across all affective tasks. As shown in Fig. 3, both training and validation losses decrease smoothly over the epochs, with the validation loss closely following the training curve, indicating the absence of overfitting and reflecting well-behaved optimization. Emotion classification accuracy exhibits a consistent upward trend, with training accuracy rising from approximately 0.79 to 0.92 and validation accuracy improving from 0.78 to about 0.88, highlighting the model's ability to capture nuanced emotional cues in multimodal Hinglish data. For the primary sentiment classification task, training accuracy steadily increases toward 0.965, whereas validation accuracy remains stable in the 0.92–0.93 range despite minor fluctuations caused by mixed-polarity samples; this suggests that the model maintains robust sentiment discrimination even under code-mixed and noisy conditions. Sarcasm detection performance is notably high, where both training and validation accuracies rapidly exceed 0.98, eventually stabilizing near 0.995 and 0.991, respectively. This near-saturated performance underscores the effectiveness of Q-OmniNet's quantum-inspired representations in capturing subtle multimodal cues associated with sarcastic expressions. Overall, the results confirm that Q-OmniNet achieves strong and stable learning behavior, delivering high predictive accuracy across emotion, sentiment, and sarcasm tasks while maintaining excellent generalization properties.

The confusion matrices in Fig. 4 provide a detailed view of Q-OmniNet's classification performance across the three affective tasks. For the seven-class emotion classification task, the model exhibits strong discriminative capability, achieving near-perfect recognition for emotions such as anger (100%), joy (99%), fear (95%), and neutral (100%). Moderate confusion appears between sadness and joy, where 22 sadness samples (76%) are correctly classified, while a small portion is misidentified as joy or surprise, indicating overlaps in linguistic or multimodal cues for these affective states. Similarly, the surprise class shows partial confusion, with 66% correctly predicted and the rest misclassified mainly as joy, reflecting semantic ambiguity typical in Hinglish social media expressions. In the sentiment classification task (three classes), the model demonstrates highly reliable performance, with negative and positive sentiments recognized with 94% and 98% accuracy, respectively. Neutral sentiment, inherently more ambiguous, achieves 40% accuracy, with misclassifications split between negative and positive due to overlapping tone and context in mixed-language posts. The sarcasm detection task achieves exceptionally high performance, with 98% accuracy for non-sarcastic and 96% for sarcastic samples. The minimal confusion—only 5 sarcastic samples classified as non-

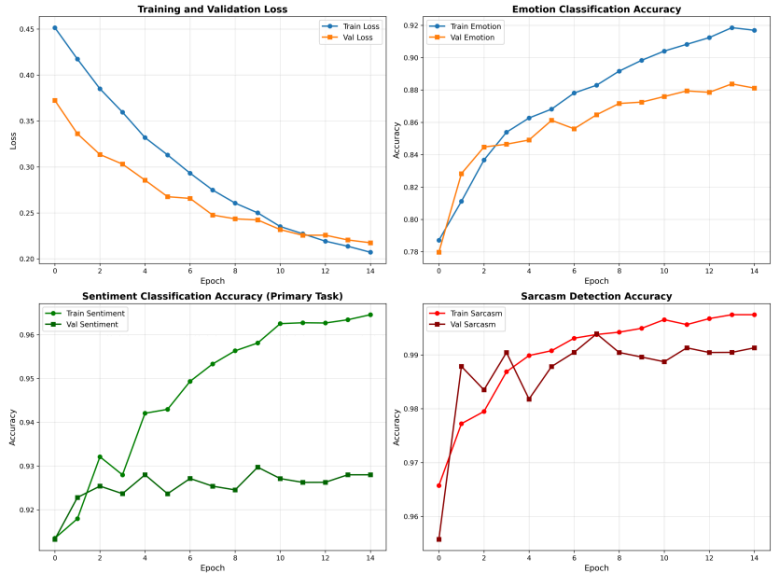

Fig. 3. Training and Validation Performance Across Emotion, Sentiment, and Sarcasm Tasks

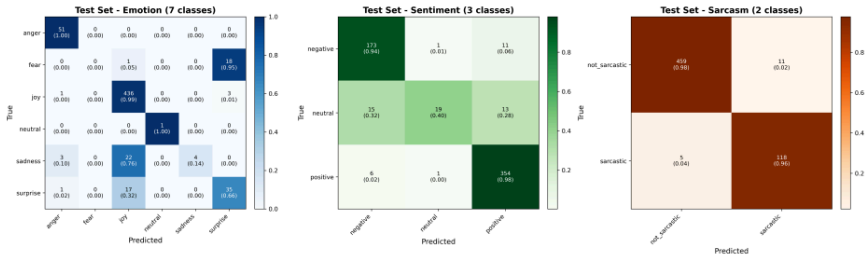

Fig. 4. Confusion Matrices for Emotion, Sentiment, and Sarcasm Classification on the Test Set

sarcastic—highlights the strength of Q-OmniNet's quantum-inspired multimodal encoding in capturing implicit and nuanced sarcastic cues. Overall, the confusion matrices affirm the model's strong generalization and its ability to distinguish fine-grained affective states despite the complexity of code-mixed Hinglish communication.

The comparative evaluation presented in Fig. 5 highlights the effectiveness of the proposed Q-OmniNet framework across three affective computing tasks. The "Test Accuracy by Task" chart shows that the classical baseline achieves moderate performance, attaining 0.74, 0.80, and 0.90 accuracy for emotion, sentiment, and sarcasm detection respectively. In contrast, the quantum-inspired variant demonstrates notable improvements across all tasks, while the quantum-hybrid model delivers the highest performance, achieving approximately 0.89 (emotion), 0.93 (sentiment), and 0.97 (sarcasm). This trend is further reflected in the "Overall Average Accuracy" plot, where the quantum-hybrid system significantly outperforms both the classical (0.7617) and quantum-inspired (0.8746) variants, reaching an average accuracy of 0.9275. The

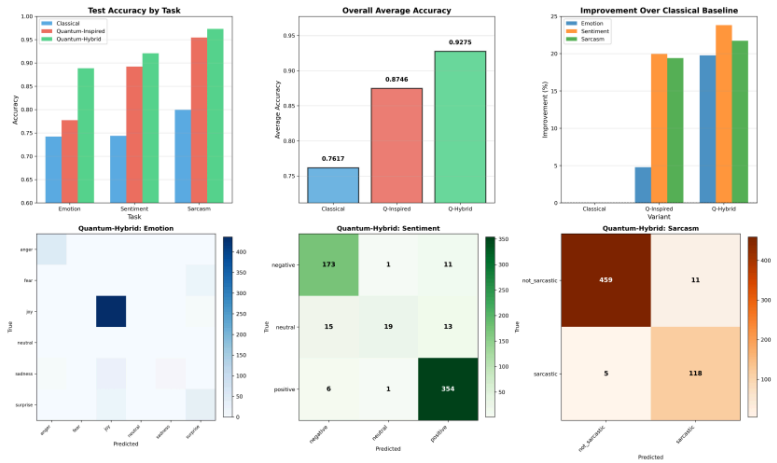

Fig. 5. Comparative Evaluation of Classical, Quantum-Inspired, and Quantum-Hybrid Models

improvement graph quantifies these gains, revealing that the quantum-inspired model offers 5–20% improvement over the classical baseline, while the quantum-hybrid variant achieves up to 25% improvement, particularly in sarcasm detection.

The confusion matrices for the quantum-hybrid model further substantiate its superior predictive capability. For emotion classification, the model achieves near-perfect recognition for high-frequency classes such as joy, while maintaining acceptable performance for lower-frequency classes such as sadness and surprise, despite the inherent ambiguity in these affective categories. Sentiment classification exhibits strong separation of polarities, with negative and positive classes predicted with high precision and recall, whereas neutral samples show moderate confusion due to overlapping contextual cues common in Hinglish social media discourse. Sarcasm detection demonstrates highly reliable performance, correctly identifying the majority of sarcastic (96%) and non-sarcastic (98%) posts, underscoring the strength of the hybrid quantum-tensor representation in capturing implicit tonal and multimodal cues. Collectively, these results validate that the quantum-hybrid design improves both discriminative power and generalization capacity across all affective tasks.

The quantum-hybrid variant achieves up to 25% improvement over the classical baseline across all tasks. Such gains can be attributed to the model's ability to represent tokens in superposed complex embedding states, enabling it to encode multiple potential meanings of polysemous Hinglish terms before context-driven disambiguation. This mechanism particularly enhances performance in emotion and sarcasm detection, where contextual polarity shifts are prevalent. Table II compares the proposed Q-OmniNet framework with existing classical and transformer-based approaches across affective classification tasks.

VIII. Conclusion and Future Work

This work presented Q-OmniNet, a quantum-inspired hybrid tensor network designed to address the intrinsic challenges of affective analysis in code-mixed and multimodal Hinglish communication. By integrating complex-valued embeddings, superposition-based semantic representations, and tensor-network–driven modeling of entangled multimodal cues, the proposed framework effectively captures ambiguity, contextual fluidity, and cross-modal dependencies that are often inadequately modeled by traditional deep learning and transformer-based approaches.

Extensive experimental evaluations demonstrate that Q-OmniNet consistently outperforms classical and transformer-based baselines across emotion, sentiment, and sarcasm classification tasks, achieving notable improvements in both accuracy and robustness. Its ability to represent multiple simultaneous meanings enables more precise handling of polysemy and context-dependent polarity shifts, which are particularly prevalent in informal, dynamic social media discourse. These results highlight Q-OmniNet's potential as a scalable and cognitively aligned paradigm for multilingual and multimodal affective computing.

Despite these advances, several avenues remain open for future research. The framework can be extended to additional multilingual and code-mixed language pairs to broaden its applicability across diverse global communication settings. Incorporating conversational history and user-level behavioral signals may further enhance the model's ability to interpret

complex, evolving emotional states. Moreover, the integration of additional modalities such as speech, audio cues, and visual information could strengthen multimodal reasoning capabilities. From a systems perspective, further optimization of the tensor network architecture and quantum-kernel components may reduce computational overhead, facilitating real-time or resource-constrained deployment. Finally, as quantum hardware continues to mature, future work may explore transitioning components of Q-OmniNet from classical simulation to actual quantum devices, potentially unlocking deeper representational advantages and paving the way toward truly quantum-native affective computing.

Q2

ORIGINALITY REPORT

4%	2%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian Institute of Technology Patna Student Paper	1%
2	www.cs.upc.edu Internet Source	<1%
3	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2022 Publication	<1%
4	Anand Babu, Saurabh G. Ghatnekar, Amit Saxena, Dipankar Mandal. "Entanglement-enabled quantum kernels for enhanced feature mapping", APL Quantum, 2025 Publication	<1%
5	Peng Zhang, Wenjie Hui, Benyou Wang, Donghao Zhao, Dawei Song, Christina Lioma, Jakob Grue Simonsen. "Complex-valued Neural Network-based Quantum Language Models", ACM Transactions on Information Systems, 2022 Publication	<1%
6	ebin.pub Internet Source	<1%
7	Submitted to Royal Holloway and Bedford New College Student Paper	<1%
8	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
9	Submitted to National Institute Of Technology, Tiruchirappalli Student Paper	<1%

10	Xun Yang, Jiaqi Zhou, Zhi Cheng, Yan Feng. "Deep learning prediction of Stokes pulse evolution in ultrafast Raman fiber amplifiers", Chinese Optics Letters, 2025 Publication	<1 %
11	publications.eai.eu Internet Source	<1 %
12	Submitted to Indian Institute of Technology, Kharagpure Student Paper	<1 %
13	Daniel Prem, Kumudha Raimond, Andrew Jeyabose. "Investigating the Impact of Feature Extraction Techniques for Classification and Synthesis of Dysarthric Speech", Circuits, Systems, and Signal Processing, 2025 Publication	<1 %
14	Jae-Hong Lee, Joon-Hyuk Chang. "Partitioning Attention Weight: Mitigating Adverse Effect of Incorrect Pseudo-labels for Self-Supervised ASR", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023 Publication	<1 %
15	Mushahid Khan, Jean Paul Latyr Faye, Udson C. Mendes, Andriy Miranskyy. "EP-PQM: Efficient Parametric Probabilistic Quantum Memory With Fewer Qubits and Gates", IEEE Transactions on Quantum Engineering, 2022 Publication	<1 %
16	Jatindra Kumar Sahu. "Introduction to Advanced Food Process Engineering", CRC Press, 2019 Publication	<1 %
17	Wesley Ferreira Maia de Souza. "End-to-End Sign Language Translation Pipeline Using Human Keypoints and Transformers", Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais, 2025	<1 %

Publication

18	Yang Wang, Xiaoxiang Liang. "Application of Reinforcement Learning Methods Combining Graph Neural Networks and Self-Attention Mechanisms in Supply Chain Route Optimization", Sensors, 2025 Publication	<1 %
19	"Az Orvosi Hetilap 1997 januári lapszámai", Orvosi Hetilap, 1997 Publication	<1 %
20	Daojun Han, Pan Qi, Juntao Zhang, Ziliang Guo, Linkun Fan. "MKDD-Vul: A Lightweight Multi-modal Knowledge Distillation Framework for Detecting Vulnerabilities in Smart Contracts", Expert Systems with Applications, 2025 Publication	<1 %
21	L. Ashok Kumar, D. Karthika Renuka. "Deep Learning Approach for Natural Language Processing, Speech, and Computer Vision - Techniques and Use Cases", CRC Press, 2023 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		