**Title:**

**Cross-Attention Transformers for Multimodal Emotion Understanding in Human-Robot Interaction**

**Abstract:**

Emotion detection is a crucial aspect of enhancing human-computer interaction by allowing machines to understand and respond to human emotional states. Two prominent methodologies have emerged in the realm of multimodal emotion recognition: a deep learning-based approach utilizing CNNs and a hybrid model combining CNNs with BERT for text processing. The CNN approach focuses on analyzing audio signals through Mel-spectrograms and MFCCs, using techniques such as Global Max Pooling and traditional machine learning algorithms (e.g., SVM, Random Forest) to classify emotions. In contrast, the hybrid model leverages the strengths of CNNs for audio feature extraction while integrating a BERT-based architecture to process textual data, achieving a significant understanding of semantic context. By employing an attention-based fusion mechanism to optimize the combination of both modalities, this model demonstrates enhanced performance, yielding an accuracy of 88.4% and an F1-score of 87.9% on the CMU-MOSEI dataset. Moreover, it achieved a weighted accuracy of 67.81% and a weighted F1 score of 66.32% on the MELD dataset. This comparison underscores the advancements in emotion detection capabilities, indicating that combining deep learning techniques with semantic comprehension leads to superior outcomes in recognizing and interpreting complex human emotions.

**Keywords:** accuracy, attention mechanism, audio processing, BERT, CNN, deep learning, emotion detection, machine learning, multimodal emotion recognition, semantic comprehension, text processing.