



XMER: An Explainable Multimodal Emotion Recognition Framework for Trustworthy Human–AI Collaboration

DOI: <https://doi.org/10.5281/zenodo.22337947>

Sneha P. Shirke¹

¹Assistant Professor Faculty of Science & Technology.

New Model Degree College Hingoli-431513, State Maharashtra, India

snehasaw1988@gmail.com

ABSTRACT

The new domain of human-agent collaboration relies more and more on emotion recognition systems that leverage speech, facial expression and text to infer affective states. The majority of the existing multimodal emotion recognition (MER) systems are predominantly black-box systems that might limit their application in critical areas like health care, education and robotic assistance. This paper proposes the combination of cross-modal attention and explainability module based on Grad-CAM, SHAP (SHapley Additive exPlanations), and attention-based modality attribution for multimodal emotion recognition in an explainable MER framework denoted XMER. Not only does XMER predict the emotion, it also provides attribution data for the input modalities and features, which enables understanding of the important factors influencing the prediction. The training goal is a combination of classification loss function and explanation-fidelity regularization. Experimental results on IEMOCAP, CMU-MOSEI and RAVDESS show that they achieve reported accuracies of 81.7% to 84.9% with higher human trust ratings. The framework offers a potential route towards more transparent and human-centered affective computing systems.

KEYWORDS: Multimodal emotion recognition; explainable artificial intelligence; XAI; SHAP; Grad-CAM; attention mechanism; trustworthy AI; human–AI collaboration; affective computing; cross-modal fusion.