ABSTRACT

Early detection of children's emotional conditions is often hindered by their limitations in direct verbal communication. Therefore, analyzing drawings and text reflections serves as an alternative psychological approach to exploring children's emotional states. However, manual assessment processes are highly subjective, while existing artificial intelligence models largely tend to overlook the depth of children's emotional intensity. This study proposes the development of a multimodal Image-Text Interaction Network architecture innovatively integrated with continuous Valence-Arousal affective features to simultaneously process the visual elements of drawings and the semantics of children's text. Experiments on the KIDO dataset show that the VA-enriched cross-modal alignment mechanism successfully reduced the relative error rate by 15.8% compared to the baseline model without Valence-Arousal, yielding a classification accuracy of 97.06% for Happiness and Sadness. Although the model is able to exploit the Valence feature as a guide, the remaining small fraction of prediction errors is primarily triggered by positively toned text that children deliberately use to mask their sadness, thereby misleading the model's sentiment. As an output, this model is functioned as an early scanning system to facilitate school counselors, homeroom teachers, and parents in providing faster, more objective, and more responsive psychological support.

***Keywords**: Affective Computing, Emotion Scanning, Multimodal, Cross-modal Fusion, Deep Learning.*

ABSTRAK

Deteksi dini kondisi emosional anak seringkali terhambat oleh keterbatasan mereka dalam berkomunikasi verbal secara langsung. Oleh karena itu, analisis coretan gambar dan teks refleksi menjadi pendekatan psikologis alternatif untuk mengeksplorasi kondisi emosional anak. Namun, proses penilaian manual bersifat sangat subjektif, sementara model kecerdasan buatan yang ada saat ini sebagian besar seringkali mengabaikan kedalaman intensitas emosi anak. Penelitian ini mengusulkan pengembangan arsitektur multimodal *Image-Text Interaction Network* yang secara inovatif diintegrasikan dengan fitur afektif kontinu *Valence-Arousal* untuk memproses elemen visual gambar dan semantik teks anak secara simultan. Eksperimen pada *dataset* KIDO menunjukan bahwa mekanisme penyelarasan *cross-modal* yang diperkaya dimensi VA berhasil menekan *error rate* relatif sebesar 15,8% dibandingkan model dasar tanpa *Valence-Arousal*, menghasilkan akurasi klasifikasi sebesar 97,06% untuk klasifikasi *Happiness* dan *Sadness*. Meskipun model mampu mengeksploitasi fitur *Valence* sebagai panduan, sisa sebagian kecil kesalahan prediksi utamanya dipicu oleh teks bernada positif yang sengaja digunakan anak untuk menutupi kesedihannya sehingga mengecoh sentimen model. Sebagai luaran, model ini difungsikan sebagai sistem pemindai dini untuk memfasilitasi Guru BK, Wali Kelas, dan Orang Tua dalam memberikan pendampingan psikologis yang lebih cepat, objektif, dan tanggap.

Kata Kunci: Komputasi Afektif, Pemindai Emosi, Multimodal, *Cross-Modal Fusion*, *Deep Learning*.