Ringkasan

Pengenalan Emosi dari Ucapan (*Speech Emotion Recognition*) memiliki potensi revolusioner dalam mengubah cara kita berinteraksi dengan teknologi, memungkinkan komputer untuk mengenali dan merespons emosi dalam ucapan manusia. Namun, keberagaman aksent dan keterbatasan dataset membuat pencapaian hasil yang optimal menjadi sulit. Penelitian ini mengeksplorasi penggunaan fitur audio mendalam untuk membantu mesin dalam membedakan emosi manusia dan menawarkan antarmuka interaksi dalam Bahasa Indonesia. Penelitian ini mengusulkan pendekatan fine-tuning berurutan (*sequential fine-tuning*) menggunakan Wav2Vec 2.0 dalam dua fase. Pada fase pertama, model dilatih awal menggunakan empat dataset SER berbahasa Inggris yang banyak digunakan: CREMA-D, RAVDESS, SAVEE, dan TESS. Pada fase kedua, model di-fine-tune lebih lanjut menggunakan dataset ucapan alami berbahasa Indonesia yang dikumpulkan dari YouTube, mencakup lima kategori emosi. Hasil penelitian menunjukkan bahwa pendekatan fine-tuning dua fase ini secara signifikan mengungguli model dasar yang hanya dilatih pada data berbahasa Indonesia, dan mampu mencapai akurasi yang mendekati performa manusia (*Human-Level Performance*). Kemudian, implementasi model disimpan di Hugging Face Space dengan Gradio untuk keperluan inferensi. Temuan ini menegaskan efektivitas metode yang diusulkan dalam meningkatkan kemampuan pengenalan emosi dalam konteks penutur Bahasa Indonesia, sehingga membuka jalan bagi interaksi manusia dan mesin yang lebih efektif.

Kata kunci: *Affective computing, Speech emotion recognition, wav2vec 2.0, speech representation, human-computer interaction.*