Pembangunan Model Deteksi Dini Gangguan Kesehatan Mental berdasarkan Ucapan: Pengenal Emosi

Riset Unggulan ITB Tahun 2026

Ahmad Mudabbir Arif , Dessi Puji Lestari, Nur Ahmadi

Pusat AI ITB

Abstract

Deteksi dini gangguan kesehatan mental melalui ucapan dan percakapan berpotensi menjadi solusi pendukung skrining awal, terutama di tengah keterbatasan tenaga profesional dan akses layanan kesehatan. Penelitian ini bertujuan mengembangkan model pembelajaran mesin berbasis ucapan untuk mengidentifikasi indikasi gangguan kesehatan mental, khususnya kecemasan.

Penelitian meliputi pembangunan korpus ucapan dan percakapan beranotasi berdasarkan instrumen psikologis terstandar, bekerja sama dengan pakar Departemen Psikiatri Universitas Padjadjaran. Fitur akustik, prosodi, emosi, serta representasi dari model pralatih ucapan dan bahasa digunakan untuk membangun model baseline. Speech Emotion Recognition (SER) menjadi salah satu komponen untuk mengenali karakteristik emosi sebagai informasi pendukung deteksi kesehatan mental. Eksperimen awal SER bahasa Indonesia menggunakan Whisper dan integrasi fitur prosodi mencapai Unweighted Accuracy 88,97%.

Penelitian diharapkan menghasilkan model dan prototipe sistem berbasis web yang akurat dan efisien untuk mendukung skrining awal kesehatan mental, bukan menggantikan diagnosis profesional. Luaran mencakup satu publikasi jurnal Q1, satu makalah konferensi internasional, dan satu prototipe sistem. Penelitian juga memperkuat kolaborasi dengan Universitas Padjadjaran dan NAIST, Jepang, serta mendukung penerapan AI for Mental Health di lingkungan pendidikan seperti ITB.

Keyword: Deteksi Dini Kesehatan Mental, Analisis Ucapan, Speech Emotion Recognition (SER), Fitur Prosodi, Whisper.

Introduction

Speech Emotion Recognition (SER) merupakan salah satu komponen pendukung deteksi kesehatan mental berbasis ucapan. Emosi yang tercermin melalui pitch, intensitas, kecepatan bicara, ritme, dan jeda dapat memberikan informasi mengenai kondisi psikologis seseorang.

Dalam penelitian ini, SER digunakan untuk mengenali karakteristik emosi sebagai salah satu indikator pendukung, yang dikombinasikan dengan fitur akustik, prosodi, dan representasi model pralatih untuk mengembangkan sistem skrining awal kesehatan mental.

Pengembangan SER memanfaatkan Whisper-base dengan adaptasi parameter efisien menggunakan LoRA dan integrasi fitur prosodi melalui acoustic prompting. Pendekatan ini dievaluasi pada korpus bahasa Indonesia dengan memperhatikan generalisasi lintas pembicara dan korpus, serta kontribusi informasi prosodi terhadap pengenalan emosi.

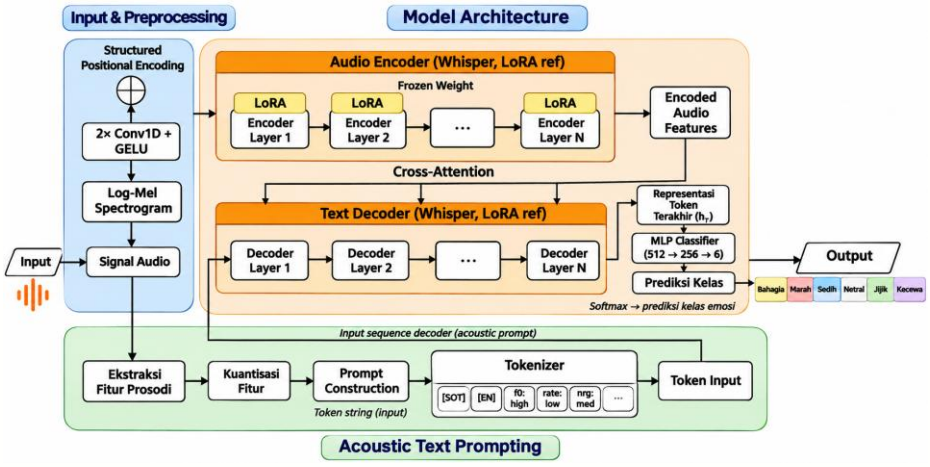
Research Method

Penelitian mengembangkan model Speech Emotion Recognition (SER) bahasa Indonesia menggunakan Whisper dan integrasi fitur prosodi melalui acoustic prompting.

Alur metode penelitian:

1. Persiapan dataset: Menggunakan 864 ujaran bahasa Indonesia dari enam kelas emosi: happy, angry, sad, neutral/question, disgust, dan disappointed.
2. Preprocessing: Audio dikonversi menjadi mono, 16 kHz, kemudian diubah menjadi log-Mel spectrogram.
3. Ekstraksi fitur prosodi: Mengekstraksi 12 fitur akustik yang mencakup pitch, energy, temporal, dan spectral.
4. Pembuatan acoustic prompt: Memilih lima fitur utama (F0 mean, F0 standard deviation, speech rate, articulation rate, dan pause ratio). Nilai fitur dikonversi menjadi token biner high/low berdasarkan median data training.
5. Pengembangan model: Whisper-base digunakan sebagai backbone. LoRA diterapkan pada encoder, sedangkan decoder dibekukan (frozen) dan menerima token prosodi melalui cross-attention.
6. Evaluasi: Membandingkan 13 konfigurasi model menggunakan Unweighted Accuracy (UA), Weighted Accuracy, Macro-F1, serta pengujian generalisasi lintas korpus.

Arsitektur SER yang diusulkan:



Gambar 1. Arsitektur SER yang Diusulkan

Discussion & Result

Model / Eksperimen	UA (%)	Hasil utama
MFCC + SVM	48,24	Baseline klasik
Wav2Vec2 Full Fine-tuning	89,72	Performa tertinggi
Wav2Vec2 + LoRA	70,28	Adaptasi parameter efisien
Whisper Audio-only (A01)	65,21	Baseline SpeechLM
Whisper + 4 fitur prosodi (B01)	85,34	Meningkat 20,13 poin dari baseline
Whisper + 5 fitur prosodi (B02)	**88,97**	**Model usulan terbaik berbasis LoRA**

Model B02 mencapai UA 88,97%, meningkat 23,76 poin dibandingkan Whisper audio-only dan mendekati Wav2Vec2 full fine-tuning (89,72%), dengan pelatihan kurang dari 1% parameter Whisper. Integrasi lima fitur prosodi melalui acoustic prompting lebih efektif dibandingkan penggunaan 12 fitur, gated fusion, maupun numeric concatenation.

Conclusion

1. Integrasi fitur prosodi pada model Whisper meningkatkan akurasi pengenalan emosi hingga 88,97%.
2. Informasi tinggi nada, kecepatan bicara, dan jeda membantu model mengenali emosi dengan lebih baik.
3. Model masih kesulitan membedakan beberapa emosi yang memiliki karakteristik suara serupa dan mengenali emosi pada korpus yang berbeda.
4. Diperlukan lebih banyak data ucapan spontan dan beragam agar model lebih akurat dan dapat digunakan pada berbagai kondisi.