I. PENDAHULUAN

1.1. Latar Belakang

Teknologi *Speech Emotion Recognition* (SER) berperan penting dalam interaksi manusia dan mesin dengan mendeteksi emosi dari sinyal suara (Akçay & Oğuz, 2020). Penerapannya meliputi kesehatan mental, layanan pelanggan, pembelajaran daring, hingga sistem keamanan berbasis suara. Namun, sebagian besar penelitian SER masih terbatas pada bahasa Inggris dan dilakukan di lingkungan terkontrol seperti laboratorium.

Di Indonesia, tantangan utama dalam pengembangan SER terletak pada keragaman bahasa dan dialek (Aji et al., 2022), keterbatasan dataset berlabel lokal, serta kurangnya infrastruktur pendukung di berbagai wilayah. Selain itu, isu privasi dan bias budaya dalam proses anotasi emosi menjadi perhatian penting dalam pengumpulan dan pengelolaan data suara. Permasalahan ini menuntut solusi yang mampu menangani keterbatasan label dan keragaman fonetik.

Beberapa pendekatan alternatif dalam pengembangan model SER mencakup penggunaan arsitektur berbasis fitur prosodik dan MFCC (*Mel-Frequency Cepstral Coefficients*). Fitur-fitur ini kemudian diklasifikasikan menggunakan model machine learning konvensional seperti SVM, Random Forest, atau MLP. Selain itu, arsitektur deep learning seperti CNN, RNN, dan LSTM juga telah banyak digunakan untuk mengekstraksi dan memproses fitur temporal dari sinyal suara (Sen et al., n.d.).

Meskipun pendekatan sebelumnya relatif lebih ringan secara komputasi, mereka memiliki keterbatasan dalam menangkap konteks global dan informasi akustik tingkat lanjut yang bersifat non-linier. Bahkan dengan model pra-latih berskala besar, performa dapat menurun secara signifikan jika dataset pada domain target terbatas dan tidak seimbang. Dalam kasus seperti ini, melakukan fine-tuning pada Wav2Vec 2.0 saja sering kali menghasilkan pembelajaran yang kurang optimal, terutama untuk kelas emosi yang kurang terwakili.

Untuk menjawab tantangan tersebut, penelitian ini mengusulkan pendekatan *sequential fine-tuning* terhadap model pra-latih Wav2Vec 2.0. Proses ini dimulai dengan pelatihan awal menggunakan dataset berbahasa Inggris berskala besar, kemudian dilanjutkan dengan *fine-tuning* kedua menggunakan dataset bahasa Indonesia yang lebih kecil. Pendekatan ini memungkinkan model mempertahankan pengetahuan umum dari pelatihan awal sekaligus menyesuaikan diri dengan karakteristik fonetik dan prosodik bahasa Indonesia.

Selain itu, pemanfaatan platform inferensi Hugging Face Hub memungkinkan implementasi model secara luas tanpa ketergantungan pada perangkat keras lokal. Dengan demikian, penelitian ini bertujuan mengembangkan model SER berbasis bahasa Indonesia yang adaptif, inklusif, dan dapat diterapkan secara praktis dalam berbagai latar belakang pengguna.

1.2. Rumusan Masalah

Dari latar belakang di atas, dirumuskan beberapa permasalahan sebagai berikut:

1. Bagaimana melatih model pengenalan emosi dalam bahasa Indonesia menggunakan *pre-trained language model* berdasarkan arsitektur *transformer*?
2. Bagaimana akurasi model pengenalan emosi terhadap bahasa Indonesia?
3. Bagaimana proses inferensi pengenalan emosi pada *Hugging Face Space*?

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Membangun model Speech Emotion Recognition (SER) berbahasa Indonesia menggunakan *pre-trained model* berbasis arsitektur Transformer.
2. Mengevaluasi akurasi model SER terhadap data suara dalam bahasa Indonesia.
3. Menerapkan model SER pada platform Hugging Face Space untuk keperluan inferensi.

1.4. Batasan Masalah

Batasan masalah dari penelitian ini adalah :

1. Penelitian ini menggunakan *Pre-trained model Wav2Vec 2.0* sebagai basis dalam pelatihan model pengenalan emosi.
2. *Dataset* yang digunakan berupa potongan audio berbahasa Indonesia yang diambil dari *podcast* di platform youtube.
3. Kategori klasifikasi berupa lima kelas emosi, yaitu: senang, sedih, marah, takut, dan jijik berdasarkan penelitian (Bryant & Barrett, 2008).

1.5. Manfaat Penelitian

Tujuan penelitian ini adalah mengeksplorasi pemanfaatan model *pre-trained Wav2Vec 2.0* berbasis transformer dalam klasifikasi emosi dari suara. Penelitian ini diharapkan menjadi referensi bagi pengembangan riset lanjutan di bidang rekognisi suara. Selain itu, penelitian ini juga bertujuan memberikan manfaat praktis dalam memahami emosi dan pengembangan aplikasi berbasis suara.