PENGENALAN EMOSI MUSIK MULTIMODAL BERBASIS LATE FUSION PADA DATASET MULTI-MODAL MIREX

PROPOSAL TUGAS BESAR
PEMBELAJARAN MESIN MULTIMODAL

Kelompok 09

Lois Novel E Gurning	122140098
Sakti Mujahid Imani	122140123
Apridian Saputra	122140143
Joshia Fernandes Sectio Purba	122140170
Sikah Nubuahtul Ilmi	122140208

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SUMATERA

2025

1. Latar Belakang dan Motivasi

Musik merupakan entitas multidimensi yang menyampaikan emosi tidak hanya melalui sinyal akustik, tetapi juga melalui semantik lirik dan struktur simbolik notasi. Studi mengenai *Music Emotion Recognition* (MER) konvensional umumnya didominasi oleh pendekatan unimodal berbasis audio. Meskipun efektif, pendekatan ini mengalami *glass ceiling effect*, di mana peningkatan akurasi mulai stagnan karena fitur audio tingkat rendah (seperti timbre dan tempo) sering kali gagal menangkap nuansa emosi yang kompleks, seperti ironi dalam lirik atau progresi akor dalam MIDI.

Analisis data eksploratif (EDA) awal pada proyek ini menunjukkan bahwa representasi fitur sederhana dari audio, lirik, dan MIDI ketika divisualisasikan menggunakan t-SNE belum membentuk klaster emosi yang terpisah secara tegas. Hal ini mengindikasikan adanya kompleksitas tinggi dan *overlapping* antar kelas emosi yang tidak dapat diselesaikan dengan model linear sederhana. Oleh karena itu, proyek ini memotivasi penggunaan arsitektur *Deep Learning* multimodal yang menggabungkan representasi fitur dari Audio (melalui CRNN), Lirik (melalui BERT), dan MIDI (melalui BiGRU) menggunakan strategi *Late Fusion* untuk menjembatani semantic gap dan meningkatkan akurasi klasifikasi.

2. Rumusan Masalah dan Tujuan Proyek

2.1 Rumusan Masalah

Berdasarkan keterbatasan pendekatan unimodal dan hasil analisis data awal, rumusan masalah dalam proyek ini adalah:

1. Bagaimana cara mengklasifikasikan emosi secara akurat menggunakan data multimodal yang terdiri dari audio, lirik, dan MIDI menggunakan metode yang paling efektif?
2. Bagaimana cara melakukan evaluasi performa model multimodal tersebut untuk menilai akurasi dari setiap modality serta hasil fusi keseluruhan?

2.2 Tujuan Proyek

Sejalan dengan rumusan masalah tersebut, tujuan dari proyek ini adalah:

1. Mengembangkan model yang tepat untuk modalitas audio, lirik, dan MIDI agar dapat mengenali emosi dengan akurat.
2. Melakukan evaluasi komprehensif terhadap model multimodal untuk mengukur akurasi dan efektivitas teknik fusion yang diterapkan.

3. Deskripsi Dataset Multimodal

3.1 Sumber dan Validitas Data

Dataset yang digunakan dalam proyek ini mengacu pada kerangka kerja *Music Emotion Recognition* (MER) multimodal yang dikembangkan oleh Panda et al. (2013). Data dikumpulkan dari basis data musik komprehensif AllMusic, yang memiliki keunggulan signifikan dibandingkan dataset musik lainnya. Berbeda dengan pendekatan *crowdsourcing* yang sering kali memiliki tingkat *noise* tinggi, label emosi pada dataset ini dikurasi oleh pakar musik profesional. Hal ini menjamin validitas dan konsistensi *ground truth* yang lebih tinggi, yang sangat krusial untuk pelatihan model *Deep Learning*.

3.2 Taksonomi Emosi (Label)

Sistem klasifikasi dalam proyek ini mengadopsi taksonomi standar dari *MIREX Mood Classification Task*. Emosi musik dikategorikan ke dalam lima klaster utama yang merepresentasikan spektrum emosi yang luas, dengan rincian sebagai berikut :

1. Cluster 1: *Passionate, rousing, confident, boisterous, rowdy*.
2. Cluster 2: *Rollicking, cheerful, fun, sweet, amiable/good natured*.
3. Cluster 3: *Literate, poignant, wistful, bittersweet, autumnal, brooding*.
4. Cluster 4: *Humorous, silly, campy, quirky, whimsical, witty, wry*.
5. Cluster 5: *Aggressive, fiery, tense/anxious, intense, volatile, visceral*.

3.3 Statistik Ketersediaan Data

Berdasarkan hasil *Exploratory Data Analysis* (EDA) yang dilakukan pada tahap awal proyek, teridentifikasi adanya ketimpangan ketersediaan data yang signifikan antar modalitas. Rincian statistik dataset mentah adalah sebagai berikut :

1. Audio: Tersedia lengkap sebanyak 903 sampel (100%) dalam format MP3 dengan durasi rata-rata 30 detik.
2. Lirik: Tersedia sebanyak 764 sampel (sekitar 85% dari total audio).
3. MIDI: Tersedia sebanyak 193 sampel (sekitar 21.5% dari total audio).

3.4 Strategi Pemilihan Data

Tantangan utama dalam arsitektur *Late Fusion* adalah kebutuhan akan input simultan dari seluruh modalitas untuk setiap sampel data. Berdasarkan temuan ketersediaan data di atas, proyek ini menerapkan strategi pemilihan data berbasis irisan (*intersection*). Hanya sampel lagu yang memiliki kelengkapan data pada ketiga modalitas (Audio, Lirik, dan MIDI) yang akan digunakan dalam proses pelatihan dan pengujian. Hal ini menghasilkan dataset

final sebanyak 193 sampel multimodal. Meskipun jumlah ini lebih kecil dibandingkan dataset unimodal audio, strategi ini memastikan integritas proses fusi dan validitas evaluasi model multimodal.

4. Rencana Metode dan Arsitektur Model

4.1 Arsitektur Model Unimodal

Proyek ini mengusulkan penerapan arsitektur *Deep Learning* multimodal untuk menangkap kompleksitas emosi musik. Setiap modalitas diproses menggunakan arsitektur pembelajaran mendalam yang dirancang khusus untuk menangkap karakteristik unik dari tipe datanya.

4.1.1 Modalitas Audio (CRNN)

Sinyal suara mentah pada modalitas audio dikonversi menjadi representasi *Mel-Spectrogram* untuk menangkap informasi frekuensi dan waktu. Arsitektur yang digunakan adalah *Convolutional Recurrent Neural Network* (CRNN) yang menggabungkan dua jenis jaringan. Lapisan *Convolutional Neural Network* berfungsi sebagai pengekstraksi fitur spasial lokal dari spektrogram seperti *timbre* dan karakteristik harmonik. Fitur tersebut kemudian diproses oleh lapisan *Recurrent Neural Network* untuk memodelkan dependensi temporal sehingga evolusi emosi sepanjang durasi lagu dapat teridentifikasi.

4.1.2 Modalitas Lirik (BERT)

Pengolahan data pada modalitas lirik melibatkan tahapan tokenisasi teks mentah dengan strategi *padding* untuk mengakomodasi variasi panjang lirik yang ditemukan pada tahap analisis data. Model yang digunakan adalah *Bidirectional Encoder Representations from Transformers* (BERT) yang telah dilatih sebelumnya. Pemilihan BERT didasarkan pada kemampuannya membaca urutan teks secara dua arah menggunakan mekanisme *attention*. Kemampuan ini krusial untuk menangkap konteks semantik yang mendalam termasuk makna ganda dalam lirik lagu yang sering terlewatkan oleh model sekuensial tradisional.

4.1.3 Modalitas MIDI (BiGRU)

Representasi data MIDI diproses sebagai urutan peristiwa musik simbolik yang mencakup atribut seperti *pitch*, *velocity*, dan durasi nada. Arsitektur yang diterapkan adalah *Bidirectional Gated Recurrent Unit* (BiGRU). BiGRU dipilih karena kemampuannya mempelajari konteks musikal dari dua arah pada data MIDI yang memiliki struktur sekuensial kuat. Unit GRU juga dinilai lebih efisien secara komputasi dibandingkan LSTM

dengan kinerja yang sebanding sehingga cocok untuk dataset MIDI yang jumlahnya terbatas.

4.2 Strategi Fusi

Integrasi fitur dilakukan pada tahap akhir menggunakan mekanisme *Late Fusion* setelah fitur diekstraksi oleh masing-masing model unimodal. Vektor probabilitas dari lapisan terakhir masing-masing *classifier* digabungkan menjadi satu representasi multimodal. Vektor gabungan tersebut kemudian diproses melalui lapisan *Fully Connected* yang berfungsi sebagai *meta-classifier* untuk mempelajari bobot kontribusi optimal dari setiap modalitas. Fungsi aktivasi *Softmax* diterapkan pada tahap akhir untuk menghasilkan probabilitas prediksi final terhadap kelima kelas emosi

5. Rencana Evaluasi dan Pembagian Peran

5.1 Rencana Evaluasi

Pengukuran kinerja sistem dilakukan secara bertingkat untuk memvalidasi kontribusi setiap modalitas sebelum digabungkan. Model unimodal yang terdiri dari audio, lirik, dan MIDI akan dievaluasi terlebih dahulu menggunakan metrik *Accuracy* dan *Confusion Matrix*. Akurasi digunakan untuk melihat *baseline* performa prediksi emosi terhadap label sebenarnya, sedangkan *Confusion Matrix* berfungsi menganalisis kesalahan klasifikasi spesifik antar kelas emosi yang berdekatan.

Evaluasi utama difokuskan pada hasil akhir sistem *Late Fusion* untuk membuktikan efektivitas penggabungan fitur. Metrik yang digunakan mencakup *Accuracy* untuk ketepatan global dan *Macro-F1 Score* sebagai indikator prioritas dalam menangani ketidakseimbangan kelas (*class imbalance*) yang ditemukan pada dataset. Analisis distribusi prediksi final juga akan dilakukan menggunakan *Confusion Matrix* untuk memverifikasi apakah kesalahan pada model unimodal berhasil diperbaiki oleh mekanisme fusi.

5.2 Pembagian Peran

Untuk memastikan pengerjaan proyek berjalan lancar dan terorganisir, berikut adalah pembagian peran dan tanggung jawab untuk setiap anggota kelompok

Tabel 1. Pembagian Peran

No	Anggota	Peran Utama	Tanggung Jawab
1.	Apridian Saputra	Pemrosesan Modalitas Lirik	- Pra-pemrosesan dan pembersihan data lirik. - Implementasi, pelatihan, dan *fine-tuning* model BERT. - Evaluasi kinerja model lirik (unimodal).
2.	Joshia Fernandes S. P	Pemrosesan Modalitas Audio	- Pra-pemrosesan data audio (konversi ke *mel-spectrogram*). - Implementasi, pelatihan, dan *tuning* model CRNN. - Evaluasi kinerja model audio (unimodal).
3.	Sikah Nubuahtul Ilmi	Pemrosesan Modalitas MIDI	- Pra-pemrosesan data MIDI (ekstraksi sekuens nada/fitur). - Implementasi, pelatihan, dan *tuning* model BIGRU. - Evaluasi kinerja model MIDI (unimodal).
4.	Lois Novel E. Gurning	Fusi Model dan Evaluasi Akhir	- Merancang dan mengimplementasikan arsitektur *late fusion*. - Mengintegrasikan output dari ketiga model. - Melatih *classifier* akhir (*FC Layer*). - Menjalankan skema evaluasi multimodal (*Accuracy*, *Macro-F1*).
5.	Sakti Mujahid Imani	Dokumentasi & Manajemen	- Penyusunan laporan proposal dan laporan akhir. - Penyiapan materi presentasi. - Memastikan integrasi kode dan sinkronisasi kemajuan tim.