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

Exploratory Data Analysis (EDA)

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. Pendahuluan

Laporan ini merangkum hasil *Exploratory Data Analysis* (EDA) yang dilakukan terhadap dataset multimodal (Audio, Lirik, dan MIDI) yang bersumber dari basis data AllMusic. Pemanfaatan dataset ini ditujukan untuk pengembangan sistem *Music Emotion Recognition* (MER) yang berfungsi mengklasifikasikan emosi musik ke dalam lima klaster emosi standar MIREX. Fase EDA ini dilaksanakan dengan tujuan utama untuk memahami karakteristik distribusi data yang mencakup analisis pada setiap modalitas secara intra-modal serta identifikasi ketersediaan data berpasangan (intersection) untuk strategi fusi inter-modal. Analisis ini selanjutnya difokuskan untuk mendeteksi potensi masalah seperti ketidakseimbangan kelas (*class imbalance*) pada target emosi dan memvalidasi separabilitas fitur awal melalui visualisasi dimensi rendah menggunakan t-SNE guna memastikan potensi klasifikasi yang optimal.

Kami telah melakukan EDA di Colab:
https://colab.research.google.com/drive/1EcaMPbq0gw4dwpmhBGRy2pr9WUhJxigf?usp=sharing#scrollTo=XAgHLbtebn4s

2. Deskripsi Dataset

2.1 Statistik Dataset

Dataset terdiri dari tiga modalitas utama dengan jumlah sampel sebagai berikut :

Tabel 1. Deskripsi Dataset

Modalitas	Format File	Jumlah Sampel	Keterangan
Audio	.mp3 / .wav	903	Data paling lengkap.
Lirik	.txt	764	~85% dari total audio.
MIDI	.mid	193	~21% dari total audio.

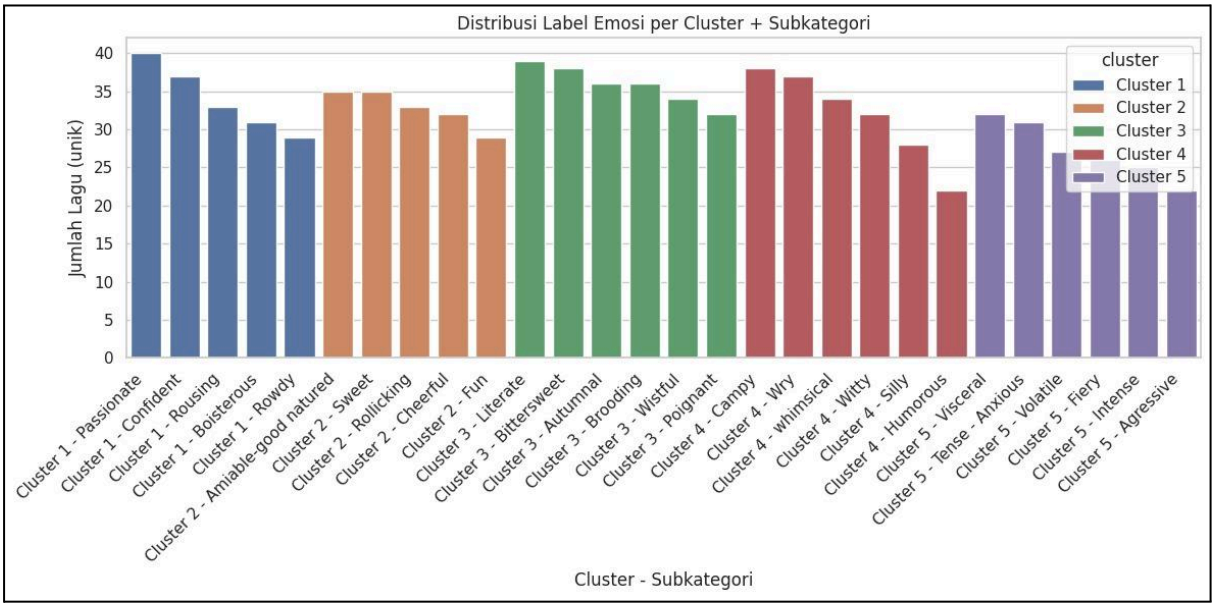
Dataset memiliki ketimpangan jumlah sampel antar modalitas yang signifikan. Perbedaan jumlah sampel antar modalitas mengindikasikan adanya tantangan *data alignment.*

2.2 Label Emosi (MIREX *Clusters*)

Data dikelompokkan menjadi 5 *cluster* emosi:

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.*

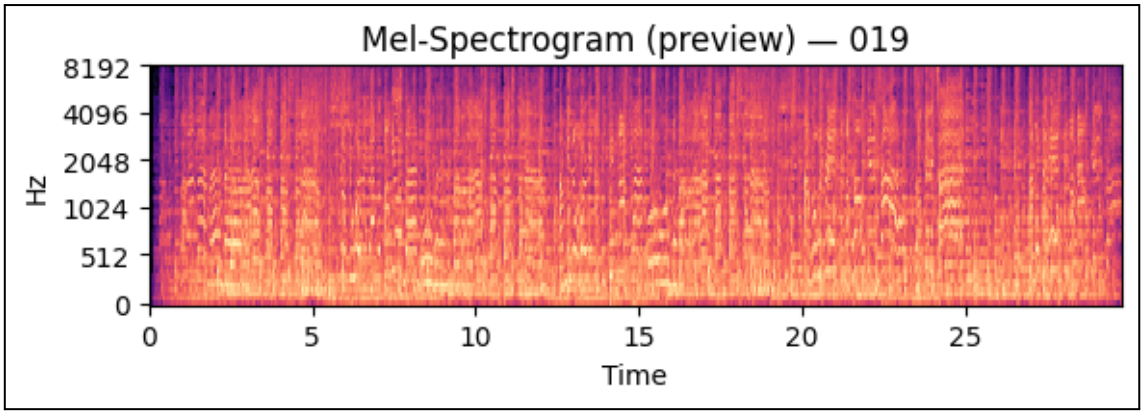


Gambar 2.1 Distribusi Label Emosi Per Cluster

3. Analisis Intra-Modal (Per Modalitas)

3.1 Modalitas Audio

Analisis dilakukan terhadap sinyal audio mentah serta representasi spektrogram guna mengekstraksi informasi akustik yang relevan. Pada tahap pemeriksaan fitur dasar, diamati adanya konsistensi durasi pada mayoritas file audio, yakni sepanjang 30 detik, yang telah sesuai dengan standar format dataset MIREX. Selanjutnya, konversi sinyal ke dalam bentuk Mel-Spectrogram diimplementasikan untuk menangkap pola distribusi frekuensi dan intensitas suara. Berdasarkan hasil visualisasi spektrogram, ditemukan perbedaan pola energi yang jelas, di mana Cluster 1 menunjukkan representasi visual yang rapat dan terang, sedangkan Cluster 3 terlihat lebih renggang dan gelap. Temuan ini mengonfirmasi bahwa representasi Mel-Spectrogram efektif dalam mendeteksi karakteristik emosi, khususnya dalam membedakan tingkat energi suara yang krusial bagi pengklasifikasian emosi musik.



Gambar 3.1 Hasil Mel-Spectogram

3.2 Modalitas Lirik

Analisis dilakukan untuk melihat sebaran panjang kata dan penggunaan kosakata. Berdasarkan visualisasi *Word Cloud*, ditemukan banyak kata yang menggambarkan perasaan kuat, seperti love, pain, dan heart, yang menunjukkan bahwa data lirik tersebut relevan untuk pengelompokan emosi. Selain itu, terlihat adanya perbedaan jumlah kata yang cukup besar antar lagu. Kondisi ini mengharuskan penggunaan teknik padding dan truncation dengan batas panjang 128 token agar data tersebut dapat diproses oleh model BERT.

Gambar 3.2 Visualisasi *Word Cloud*

Top 5 Bigram per Label

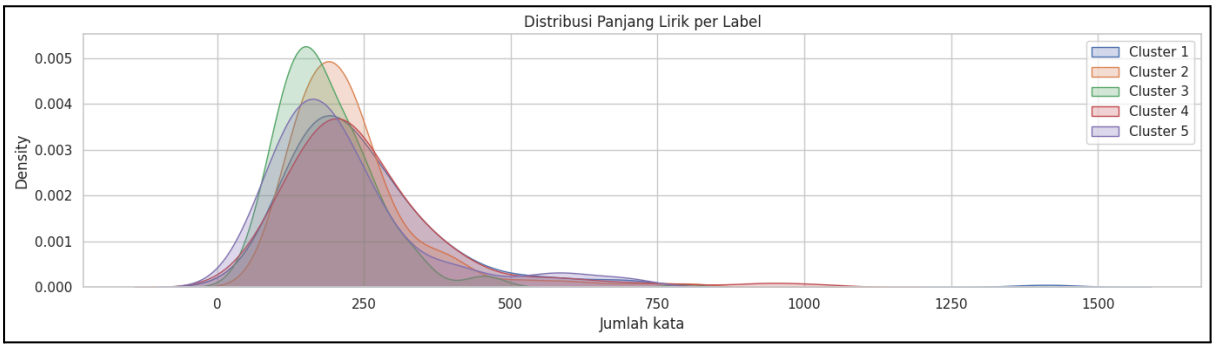
[Cluster 1] : *love cause, tight tonight, said come, heard news, know like*

[Cluster 2] : *cause im, yeah im, youve got, like im, hes got*

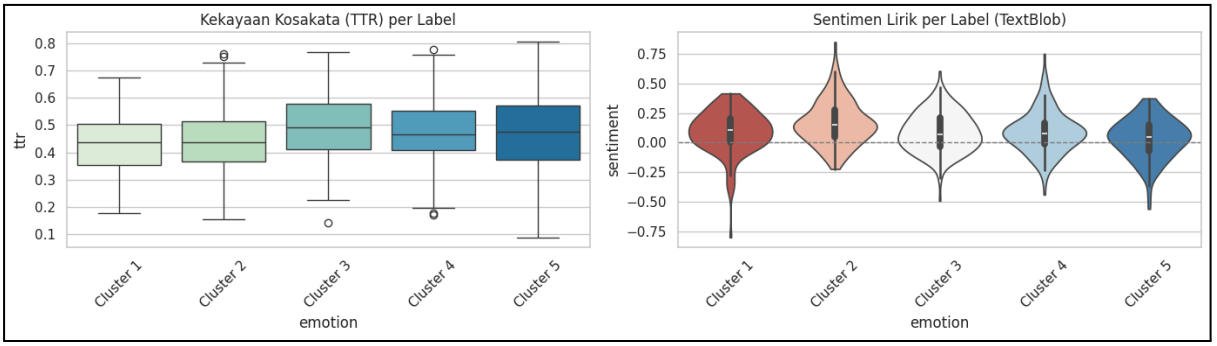
[Cluster 3] : *come true, things said, want know, years ago, love love*

[Cluster 4] : *dont like, speak said, didnt like, tell ya, man oh*

[Cluster 5] : *tell im, yeah know, got dont, old man, know dont*



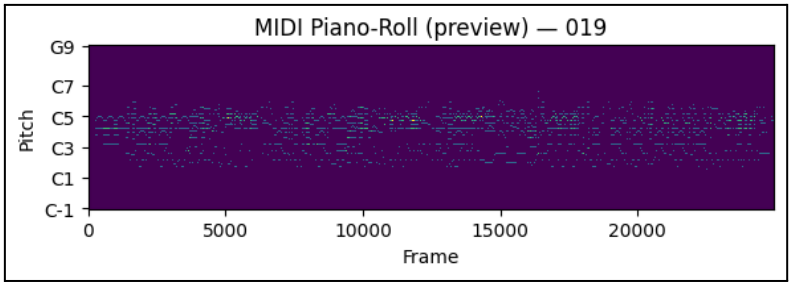
Gambar 3.3 Distribusi Panjang Lirik



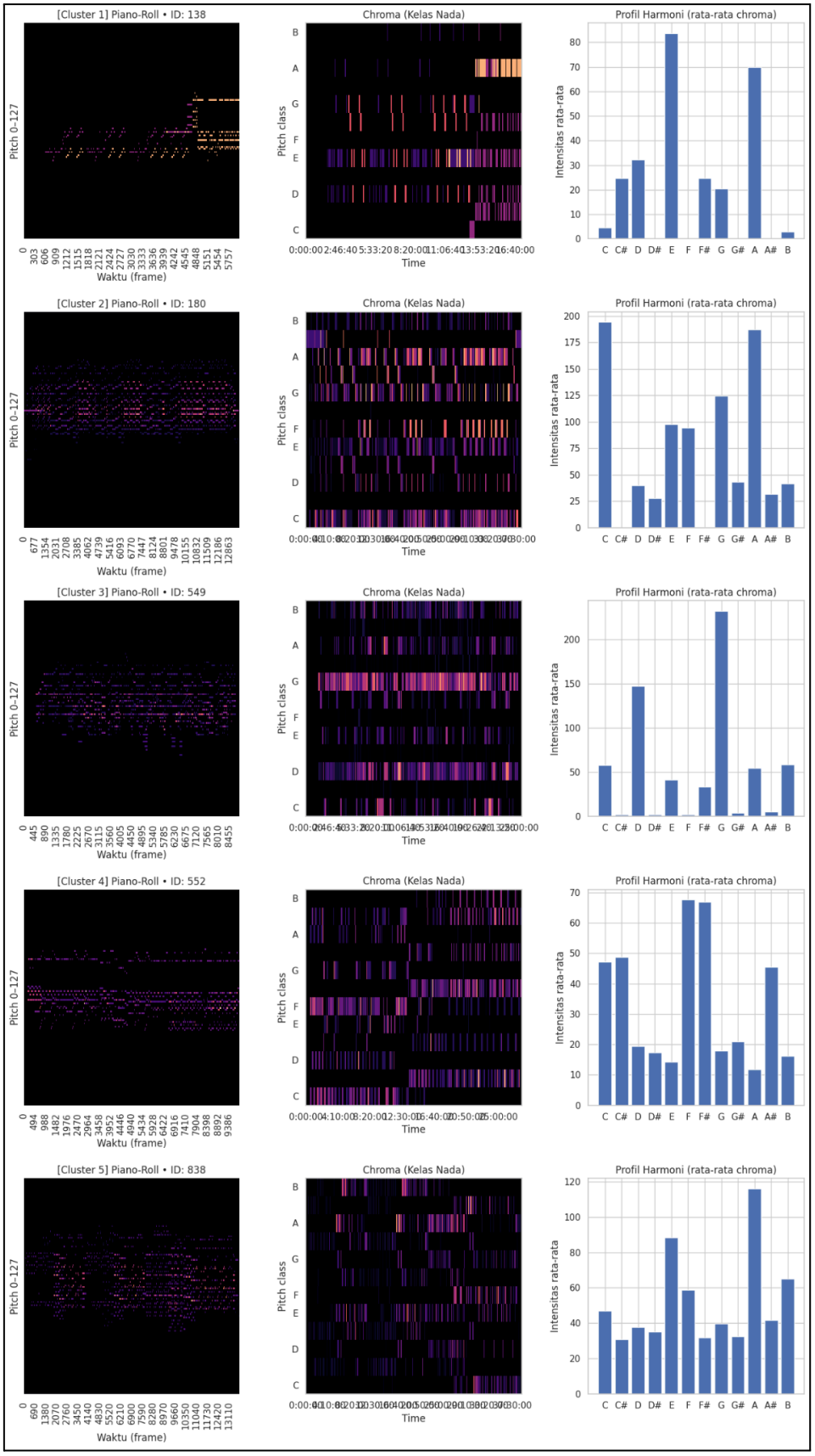
Gambar 3.4 Sentimen Lirik

3.3 Modalitas MIDI

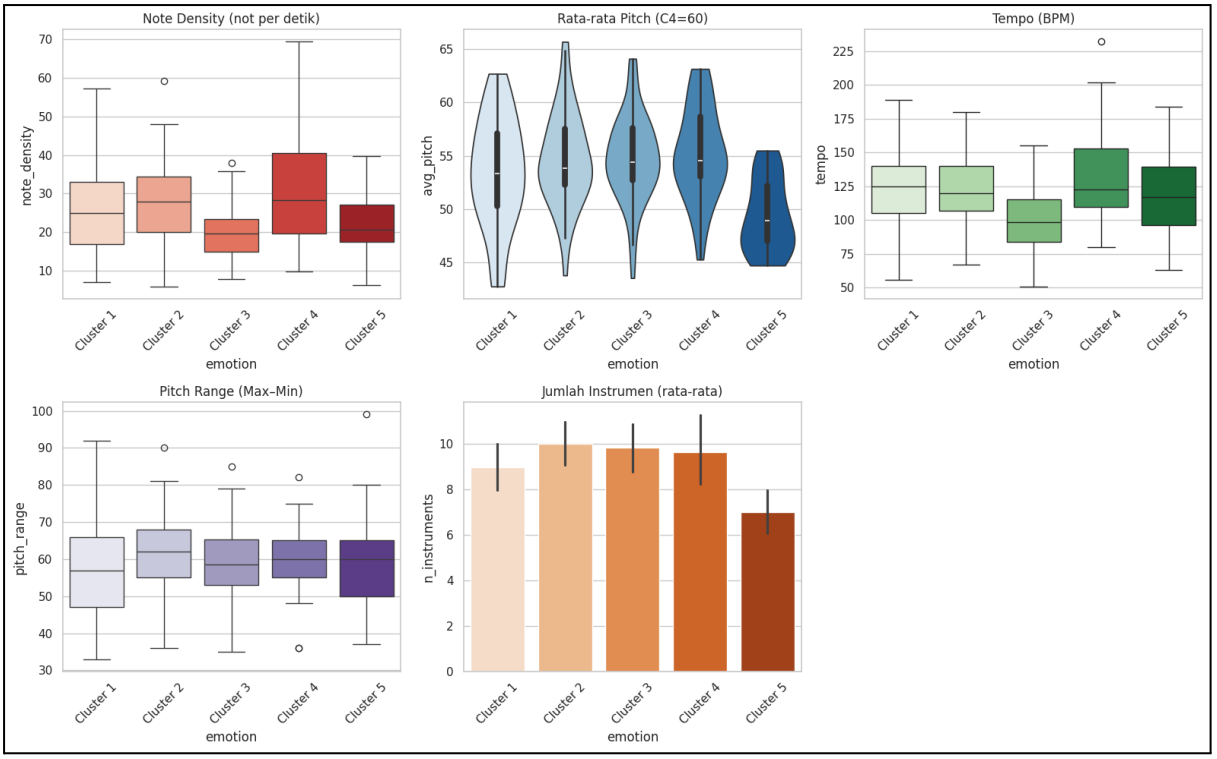
Analisis terhadap data MIDI difokuskan pada distribusi pitch dan velocity. Data ini direpresentasikan sebagai rangkaian note events, di mana visualisasi Piano Roll memperlihatkan perbedaan kerapatan nada yang jelas antar emosi. Selain itu, ditemukan struktur kombinasi dua atau lebih nada atau baris melodi yang dimainkan secara bersamaan yang kompleks, khususnya pada lagu dengan tempo cepat yang memiliki kepadatan aktivitas nada (*event density*) lebih tinggi. Jumlah sampel MIDI yang tersedia juga jauh lebih sedikit dibandingkan dengan data audio dan lirik, yakni hanya 193 sampel. Keterbatasan jumlah data ini menjadi kendala utama dalam penerapan pelatihan model fully-multimodal pada seluruh dataset.



Gambar 3.4 Visualisasi Piano-Roll



Gambar 3.5 Visualisasi MIDI



Gambar 3.6 Hasil EDA MIDI

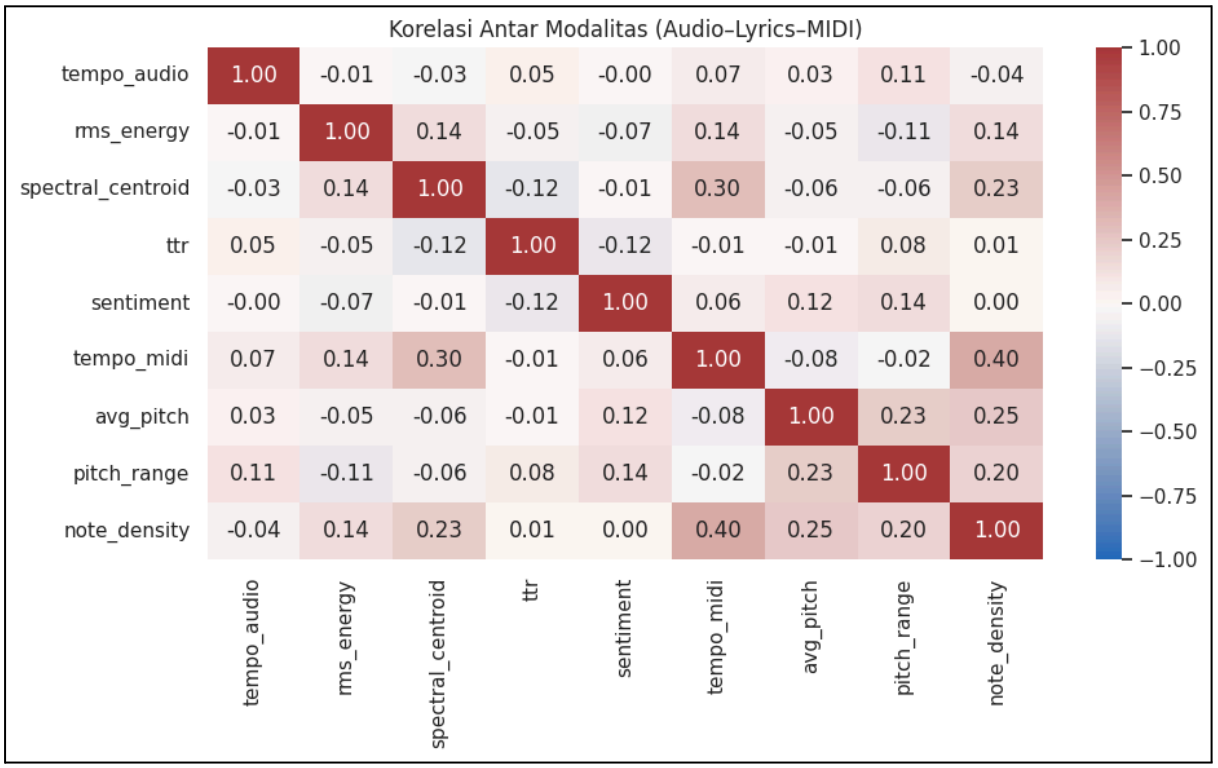
4. Analisis Inter-Modal (Korelasi Antar Modalitas)

Fokus utama analisis *inter-modal* ini adalah meninjau ketersediaan data berpasangan guna menentukan strategi fusi yang tepat. Tantangan terbesar yang ditemukan adalah ketidaksamaan jumlah sampel antar modalitas atau *missing modalities*. Berdasarkan data yang ada, Audio memiliki sekitar 903 sampel dan Lirik 764 sampel, sedangkan MIDI hanya tersedia sebanyak 193 sampel. Kondisi ini menunjukkan bahwa hanya sebagian kecil *dataset* yang memiliki ketiga modalitas secara lengkap. Jika metode *Early Fusion* diterapkan, jumlah data latih akan terpangkas drastis mengikuti jumlah sampel MIDI yang sedikit. Oleh karena itu, penerapan arsitektur *Late Fusion* menjadi keputusan desain yang paling tepat. Dengan strategi ini, model Audio, Lirik, dan MIDI dapat dilatih secara independen memaksimalkan data yang tersedia, dan penggabungan dilakukan pada tahap prediksi akhir. Selain itu, ditemukan pula tantangan teknis lain seperti perlunya pencocokan ID *file* (*alignment*) yang akurat, ketidakseimbangan kelas yang didominasi oleh *Cluster* 1 dan 2, serta adanya *noise* atau karakter sampah pada teks lirik yang memerlukan pembersihan lebih lanjut.

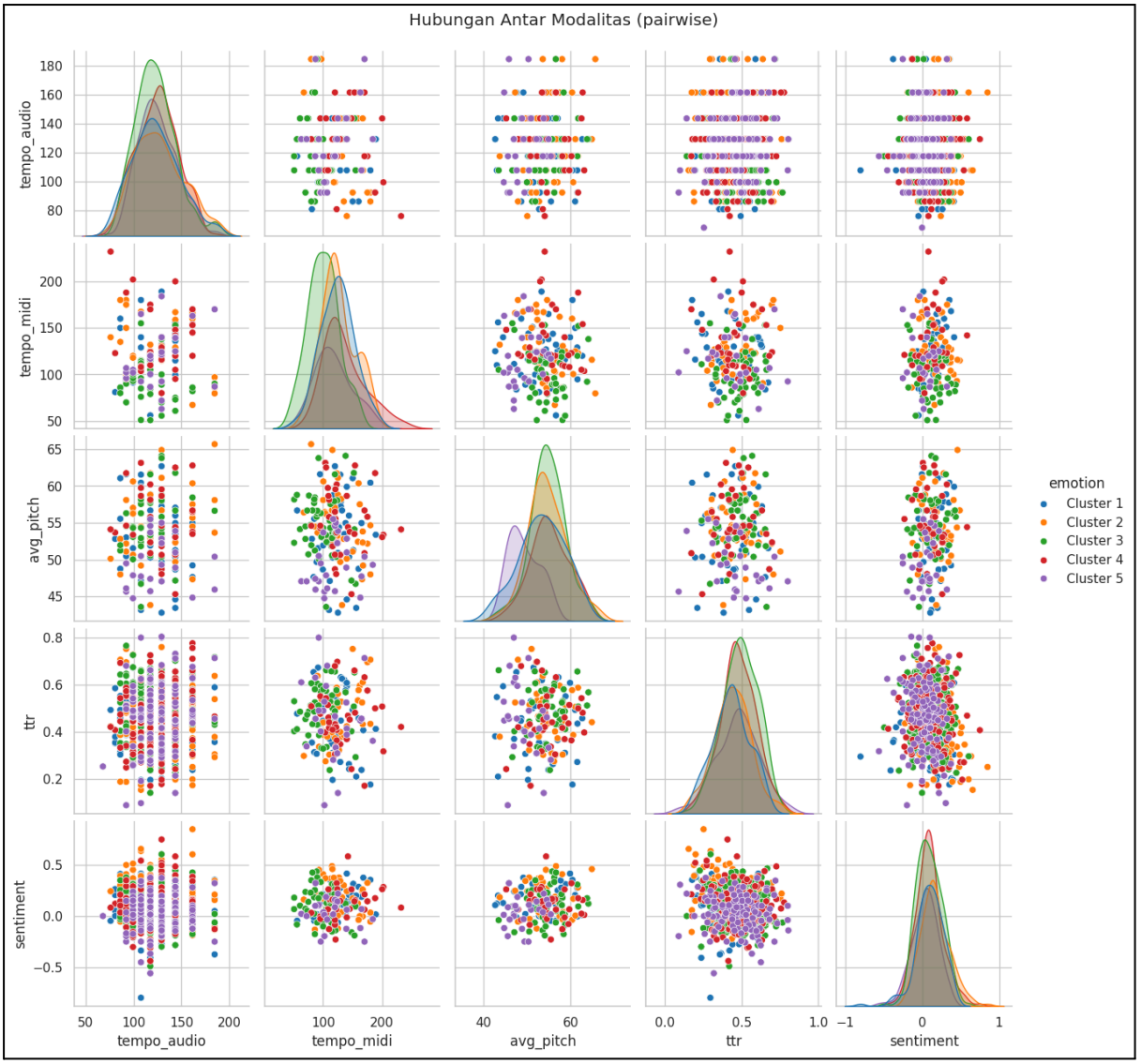
ALIGNMENT ANTAR MODALITAS (berdasarkan file_id_norm)

Total ID unik yang muncul di ≥1 modalitas: 903

1. ID yang TIDAK punya audio (*missing* audio): 0
2. ID yang TIDAK punya lyrics (*missing* lyrics): 139
3. ID yang TIDAK punya midi (*missing* MIDI): 709



Gambar 4.1 Korelasi Antar Modalitas



Gambar 4.2 Visualisasi Hubungan Antar Modalitas

5. Analisis Target

Analisis mendalam terhadap variabel target mengungkapkan adanya ketidakseimbangan distribusi kelas (*class imbalance*) yang cukup signifikan dalam dataset, yang berpotensi mempengaruhi kinerja model klasifikasi. Berdasarkan perhitungan frekuensi label, ditemukan bahwa kelas emosi tertentu seperti *Passionate* dan *Literate* mendominasi jumlah sampel, sementara kelas lain seperti *Humorous* dan *Aggressive* memiliki representasi yang jauh lebih sedikit. Ketimpangan ini menjadi semakin ekstrem ketika analisis dipersempit hanya pada subset data yang memiliki modalitas lengkap (Audio, Lirik, dan MIDI), di mana beberapa label emosi hanya diwakili oleh beberapa sampel (kurang dari 5 data). Jika tidak ditangani dengan strategi seperti *stratified splitting* atau pembobotan kelas (*class weighting*), model berisiko mengalami bias dengan cenderung memprediksi kelas mayoritas dan mengabaikan nuansa emosi pada kelas minoritas

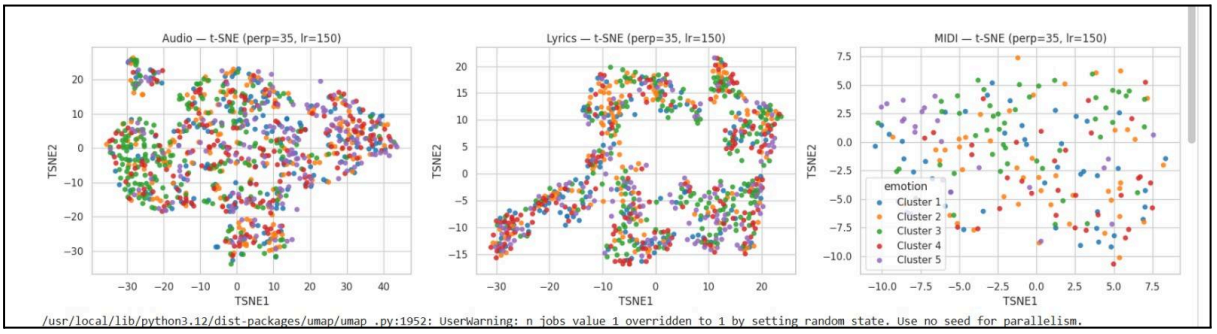
Dalam upaya menghubungkan temuan fitur dengan label tujuan akhir, visualisasi ruang fitur menggunakan PCA dan t-SNE menunjukkan tantangan separabilitas yang nyata. Secara umum, representasi fitur mentah (*embedding*) dari ketiga modalitas belum mampu membentuk klaster emosi yang terpisah secara tegas; sebaliknya, titik-titik data antar kelas terlihat saling tumpang tindih (*overlap*) secara signifikan dalam ruang dua dimensi. Hal ini mengindikasikan bahwa korelasi antara fitur fisik (seperti spektrum suara atau urutan nada) dengan label emosi abstrak bersifat sangat kompleks dan non-linear. Temuan ini menegaskan bahwa penggunaan fitur mentah secara langsung tidak cukup untuk membedakan target emosi, sehingga validasi penggunaan arsitektur *Deep Learning* yang mampu mempelajari representasi non-linear menjadi mutlak diperlukan.

Terkait identifikasi kekuatan pola per modalitas, analisis visual memperlihatkan karakteristik yang berbeda-beda dalam mengindikasikan label tertentu. Modalitas teks (lirik) menunjukkan struktur pengelompokan yang relatif lebih teratur dibandingkan modalitas lainnya, menyiratkan bahwa konten semantik lirik memiliki korelasi yang lebih kuat dan langsung terhadap label emosi dibandingkan fitur lainnya. Di sisi lain, modalitas audio dan MIDI memperlihatkan pola penyebaran yang jauh lebih acak dan tersebar luas, yang menandakan bahwa fitur-fitur tersebut pada tahap ini lebih dominan menangkap karakteristik teknis (seperti akustik instrumen atau struktur notasi) daripada ekspresi emosionalnya. Oleh karena itu, penggabungan ketiga modalitas melalui mekanisme fusi (*Late Fusion*) diharapkan dapat saling melengkapi kekurangan masing-masing modalitas untuk memperkuat indikasi terhadap label target.

6. Analisis Visualisasi Fitur

6.1 t-SNE

Teknik t-SNE (*t-Distributed Stochastic Neighbor Embedding*) digunakan untuk memproyeksikan fitur berdimensi tinggi dari audio dan modalitas lain ke dalam ruang 2D. Visualisasi menggunakan t-SNE dilakukan pada masing-masing modalitas untuk melihat keterhubungan antar modalitas. Sebaran dari data ini dapat dilihat pada Gambar 6.1 di bawah ini.



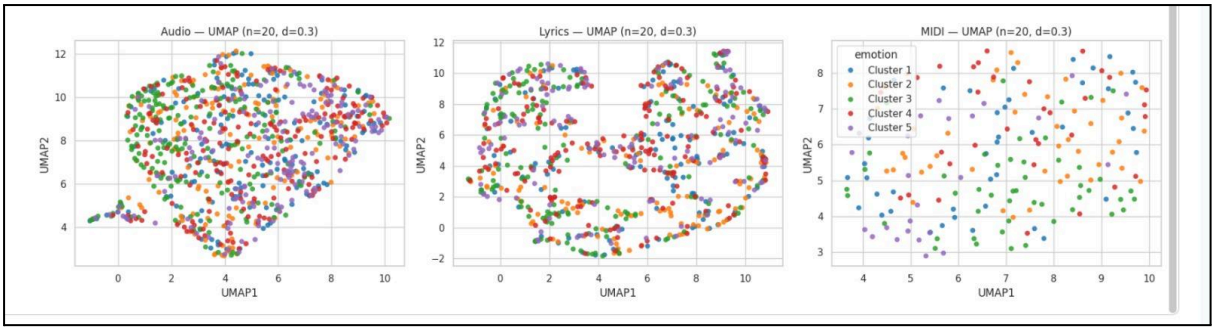
Gambar 6.1 Visualisasi t-SNE

Sebaran data pada modalitas audio menunjukkan pola yang relatif lebih terstruktur, meskipun klaster masih saling tumpang tindih. Karakteristik ini muncul karena fitur audio memiliki korelasi lebih langsung dengan ekspresi emosi, seperti tempo, loudness, dan timbre. Sebaran t-SNE untuk modalitas lirik terlihat lebih menggerombol namun tidak menciptakan formasi klaster berdasarkan label emosi. Pola ini wajar karena fitur semantik dasar seperti polaritas sentimen dan TTR tidak cukup kaya untuk membedakan kategori emosional secara eksplisit. Sebaran t-SNE untuk modalitas MIDI memperlihatkan struktur yang paling acak dan menyebar. Ketidakstabilan ini muncul karena representasi MIDI sangat dipengaruhi variasi instrumentasi, densitas not, dan pitch pattern yang tidak selalu berkaitan langsung dengan label emosi.

Plot t-SNE menunjukkan adanya pengelompokan awal berdasarkan label emosi, meskipun masih terdapat *overlap* antar kelas yang berdekatan . Hal ini menandakan bahwa fitur yang diekstrak mengandung informasi diskriminatif yang cukup baik, namun memerlukan model non-linear seperti *Deep Learning* untuk memisahkan batas keputusan (*decision boundary*) dengan lebih akurat.

6.2 UMAP

Teknik *Uniform Manifold Approximation and Projection* (UMAP) diterapkan untuk memproyeksikan fitur berdimensi tinggi dari audio dan modalitas lainnya ke dalam ruang 2D.

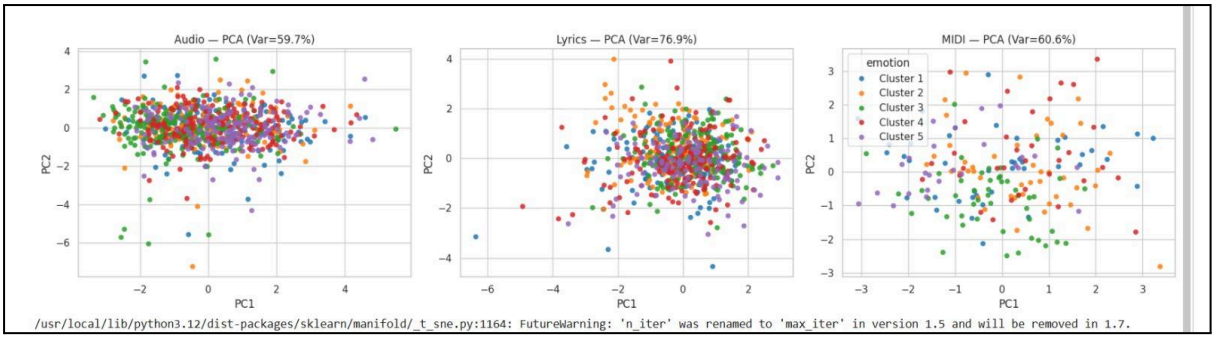


Gambar 6.2 Visualisasi UMAP

Berdasarkan analisis data gabungan yang mencakup audio, lirik, dan MIDI, dapat disimpulkan bahwa fitur-fitur yang diekstrak memiliki informasi yang kaya untuk mengenali emosi lagu, sebagaimana terlihat dari terbentuknya pengelompokan awal pada visualisasi UMAP. Namun, karena masih terdapat tumpang tindih (*overlap*) yang signifikan antar-kategori emosi yang menunjukkan bahwa batas pemisahnya tidak sederhana atau linear, maka metode pemrosesan biasa tidaklah cukup. Oleh karena itu, pengembangan proyek ini mutlak memerlukan penerapan metode Deep Learning yang lebih canggih untuk mengurai pola rumit tersebut dan meningkatkan akurasi klasifikasi emosi secara signifikan

6.3 PCA

Teknik PCA (Principal Component Analysis) digunakan untuk mereduksi dimensi fitur dengan cara memproyeksikannya ke dalam komponen yang mampu menjelaskan variansi terbesar dari data. Pendekatan ini bekerja dengan mencari kombinasi linear dari fitur asli sehingga struktur global dataset dapat diamati secara lebih sederhana. Proyeksi dua dimensi pada PCA memberikan gambaran mengenai apakah kelas emosi memiliki separabilitas linear yang cukup untuk dipetakan tanpa transformasi non-linear tambahan. Hasil visualisasi menunjukkan bahwa persebaran data masih saling bertumpang tindih pada ruang komponen utama, sehingga struktur linear tidak mampu memisahkan kelima klaster emosi secara jelas. Informasi ini mengindikasikan bahwa hubungan antar fitur multimodal bersifat non-linear serta memerlukan pendekatan pembelajaran representasi yang lebih dalam agar pola emosional dapat ditangkap secara lebih akurat.



Gambar 6.3 Visualisasi PCA

Plot PCA untuk modalitas audio memperlihatkan persebaran horizontal yang lebar, dipengaruhi oleh variasi tempo, spectral centroid, dan RMS antar lagu. Meski demikian, klaster emosi tetap saling tumpang tindih. PCA untuk modalitas lirik menunjukkan variansi terjelaskan terbesar (±76.9%) sehingga dimensi utamanya mampu menangkap sebagian besar variasi teks. Persebaran titik tetap tidak membentuk klaster emosional yang jelas karena fitur permukaan seperti TTR dan sentimen tidak cukup representatif. PCA untuk modalitas MIDI menampilkan persebaran yang cukup heterogen dengan variasi pitch range dan note density sebagai sumber variansi dominan. Klaster emosional tetap tidak terpisah sehingga struktur linear pada fitur MIDI juga belum cukup diskriminatif.