Jurnal Riset Aplikasi Ilmu Data dan Sistem 1 (1), 47–61 DOI: 0000
Juli 2026

Pengenalan Emosi Berbasis Multimodal Data Menggunakan Pendekatan Ensemble Deep Learning

Fadhilah Nuria Shinta[1], Elly Matul Imah[2,*]

[1]*Program Studi S1 Sains Data, Universitas Negeri Surabaya, Indonesia*
[2]*Program Studi S1 Kecerdasan Artifisial, Universitas Negeri Surabaya, Indonesia*

Abstrak

Pengenalan emosi menjadi salah satu aspek penting dalam pengembangan sistem interaksi manusia dan komputer yang mampu beradaptasi dan merespons kondisi penggunanya. Penelitian ini bertujuan untuk mengembangkan model klasifikasi emosi berbasis multimodal dengan memanfaatkan data teks dan audio dari dataset MELD. Pada modalitas teks digunakan model berbasis transformer, yaitu DistilBERT, sedangkan pada modalitas audio digunakan kombinasi Convolutional Neural Network (CNN) dan Bidirectional Gated Recurrent Unit (BiGRU). Penggabungan kedua modalitas dilakukan menggunakan metode *ensemble* berupa *soft voting* dan *harmonic mean*. Hasil penelitian menunjukkan bahwa model teks memberikan performa terbaik pada pendekatan unimodal dengan akurasi sebesar 63%, Macro F1-score (MF1) sebesar 45,3%, dan Weighted F1-score (WF1) sebesar 62%, sedangkan model audio memperoleh akurasi sebesar 46%, MF1 sebesar 14,4%, dan WF1 sebesar 35%. Pada pendekatan multimodal, metode *soft voting* menghasilkan akurasi sebesar 64%, MF1 sebesar 45,3%, dan WF1 sebesar 62%, sementara metode *harmonic mean* menghasilkan akurasi terbaik sebesar 65%, MF1 sebesar 44,6%, dan WF1 sebesar 62%. Hasil tersebut menunjukkan bahwa informasi dari data teks memberikan kontribusi yang lebih besar dibandingkan informasi data audio dalam klasifikasi emosi pada dataset MELD. Meskipun pendekatan multimodal meningkatkan akurasi, nilai MF1 dan WF1 belum menunjukkan peningkatan dibandingkan model teks. Secara keseluruhan, penelitian ini menunjukkan bahwa metode ensemble sederhana tetap mampu memberikan performa yang kompetitif dalam klasifikasi emosi multimodal.

Kata Kunci *CNN-BiGRU; DistilBERT; harmonic mean; multimodal; MELD; pengenalan emosi; soft voting*

Abstract

Emotion recognition is one of the essential aspects in developing human-computer interaction systems that can adapt to and respond to users' conditions. This study aims to develop a multimodal emotion classification model by utilizing text and audio data from the MELD dataset. For the text modality, a transformer-based model, DistilBERT, was employed, while the audio modality utilized a combination of Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (BiGRU). The two modalities were integrated using *ensemble* methods, namely *soft voting* and *harmonic mean*. The experimental results show that the text model achieved the best performance in the unimodal approach, with an accuracy of 63%, a Macro F1-score (MF1) of 45.3%, and a Weighted F1-score (WF1) of 62%. Meanwhile, the audio model obtained

*Corresponding author. Email: ellymatul@unesa.ac.id.

Received 12 Juni 2026; Revised 10 Juli 2026; Accepted 12 Juli 2026

an accuracy of 46%, an MF1 of 14.4%, and a WF1 of 35%. In the multimodal approach, the *soft voting* method achieved an accuracy of 64%, an MF1 of 45.3%, and a WF1 of 62%, while the *harmonic mean* method achieved the best accuracy of 65%, an MF1 of 44.6%, and a WF1 of 62%. These results indicate that textual information contributes more significantly than acoustic information to emotion classification on the MELD dataset. Although the multimodal approach improved classification accuracy, the MF1 and WF1 values did not show improvement compared to the text model. Overall, this study demonstrates that simple *ensemble* methods are still capable of delivering competitive performance for multimodal emotion classification.

Keywords *CNN-BiGRU; DistilBERT; harmonic mean; multimodal; MELD; emotion recognition; soft voting*

1 Pendahuluan

Emosi merupakan bagian mendasar dalam kehidupan manusia yang memengaruhi cara berpikir, berinteraksi, mengambil keputusan, serta beradaptasi terhadap lingkungan. Bagi siswa, kemampuan mengenali dan mengelola emosi menjadi keterampilan penting yang berkontribusi terhadap proses pembelajaran, hubungan sosial, dan perkembangan karakter. Berbagai studi menunjukkan bahwa kemampuan regulasi emosi siswa dipengaruhi oleh lingkungan sekolah. Penelitian di Indonesia melibatkan lebih dari 13 ribu remaja dan menunjukkan adanya hubungan signifikan antara iklim sekolah dan regulasi emosi (Dadeh et al., 2025). Siswa yang merasa aman dan memiliki keterikatan positif terhadap lingkungan sekolah cenderung memiliki kemampuan regulasi emosi yang lebih baik. Selain itu, kecerdasan emosional juga berhubungan positif dengan prestasi akademik dan motivasi belajar (Tang and He, 2023). Hal ini menunjukkan bahwa dukungan terhadap pemahaman emosi memiliki peran penting dalam proses pendidikan.

Di sisi lain, perkembangan teknologi juga membuka peluang baru untuk membantu proses pengenalan dan pengelolaan emosi. Harahap et al. (2024) mengembangkan sistem EMOKIDS untuk mendeteksi ekspresi wajah anak sekolah dasar. Sistem tersebut mampu mengenali emosi dasar seperti senang, marah, sedih, takut, dan cemas. Hasil ini menunjukkan bahwa teknologi dapat digunakan untuk mendukung pemantauan kondisi emosional siswa dalam proses pembelajaran. Selain itu, pada kompetisi Speech Emotion Recognition di Interspeech 2025, terdapat model berbasis suara yang mampu meningkatkan performa prediksi emosi dengan mempertimbangkan karakteristik pembicara serta analisis vokal (Tang, 2025).

Ekspresi wajah merupakan salah satu modalitas utama dalam pengenalan emosi karena merepresentasikan aktivitas otot yang mencerminkan kondisi emosional individu (Ekman, 1992; Martinez, 2017). Modalitas ini menghadapi berbagai tantangan seperti variasi individu, perbedaan pencahayaan, sudut pengambilan gambar, serta keterbatasan data berlabel (Jayaswal et al., 2025). Selain ekspresi wajah, emosi juga dapat dikenali melalui teks yang umumnya diekspresikan secara implisit melalui pilihan kata dan struktur kalimat, meskipun sering terkendala oleh ambiguitas kalimat. Sementara itu, suara menjadi modalitas penting lainnya, di mana emosi tercermin dari intonasi, prosodi, dan ritme bicara, namun tetap rentan terhadap noise serta variasi bahasa (Larrouy-Maestri et al., 2024; Jordan et al., 2025). Untuk mengatasi keterbatasan masing-masing modalitas, pendekatan multimodal digunakan dengan menggabungkan informasi dari teks, suara, dan visual sehingga menghasilkan representasi emosi yang lebih lengkap. Salah satu penerapannya adalah sistem MIMOSYS yang memanfaatkan analisis suara untuk pemantauan kondisi mental dan telah digunakan dalam konteks nyata seperti penanganan pascabencana (Higuchi et al., 2020).

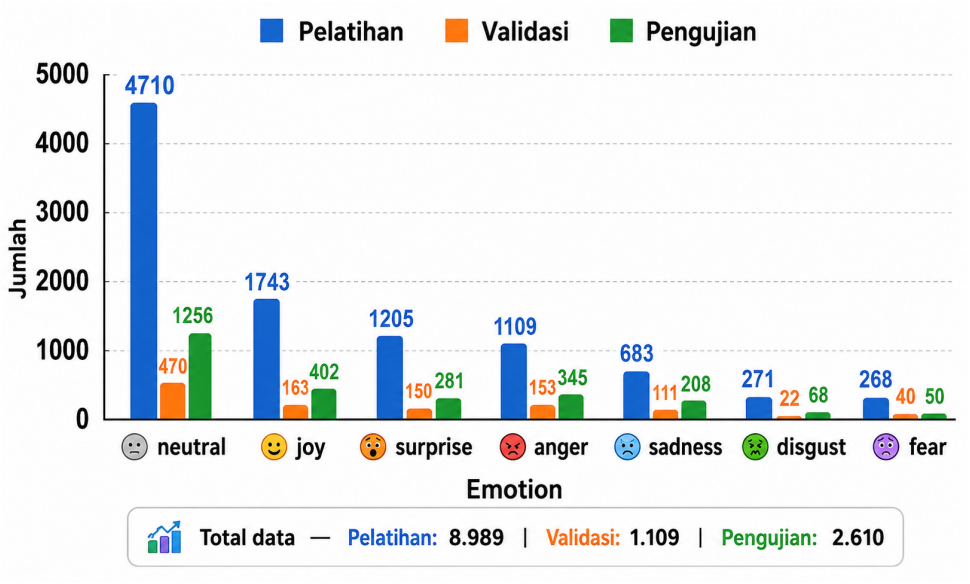
Berdasarkan hal tersebut, penelitian ini mengembangkan model pengenalan emosi multimodal berbasis teks dan audio menggunakan dataset MELD (Poria et al., 2019). Modalitas teks diproses menggunakan DistilBERT, sedangkan modalitas audio menggunakan kombinasi CNN dan BiGRU untuk mempelajari karakteristik spasial dan temporal sinyal suara. Selanjutnya, hasil prediksi kedua modalitas digabungkan menggunakan metode ensemble berupa *soft voting* dan *harmonic mean* untuk menghasilkan keputusan klasifikasi emosi akhir.

Kontribusi utama penelitian ini terletak pada pengembangan pendekatan multimodal berbasis ensemble sederhana yang mengombinasikan model transformer pada data teks dan data audio. Berbeda dengan beberapa penelitian sebelumnya yang menggunakan arsitektur *fusion* kompleks, penelitian ini mengevaluasi efektivitas metode *soft voting* dan *harmonic mean* dalam menghasilkan performa klasifikasi emosi yang lebih stabil dengan kompleksitas yang lebih rendah. Selain itu, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan kajian *multimodal emotion recognition*, khususnya pada pemanfaatan dataset MELD.

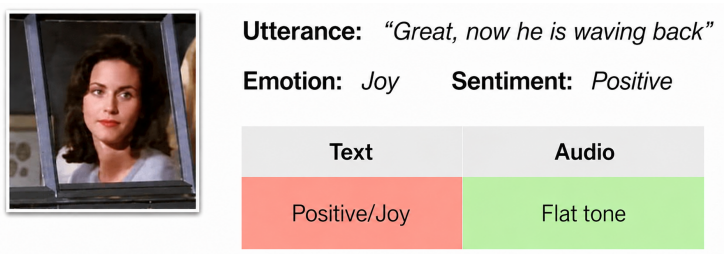
2 Metode

2.1 Dataset

Penelitian ini menggunakan *Multimodal EmotionLines Dataset* (MELD) yang dikembangkan dari serial televisi *Friends*. Dataset ini terdiri atas 1.433 percakapan dengan total 13.708 ujaran yang telah dianotasi ke dalam tujuh kategori emosi, yaitu *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, dan *surprise* (Poria et al., 2019). Setiap ujaran dilengkapi dengan transkrip teks, identitas pembicara, serta data audio dan visual. Pada penelitian ini, hanya modalitas teks dan audio yang digunakan sebagai masukan model. Pembagian data mengikuti konfigurasi resmi MELD yang terdiri atas data pelatihan, validasi, dan pengujian. Rincian jumlah percakapan dan ujaran pada masing-masing subset ditunjukkan pada Gambar 1. Contoh sampel data yang digunakan ditunjukkan pada Gambar 2.



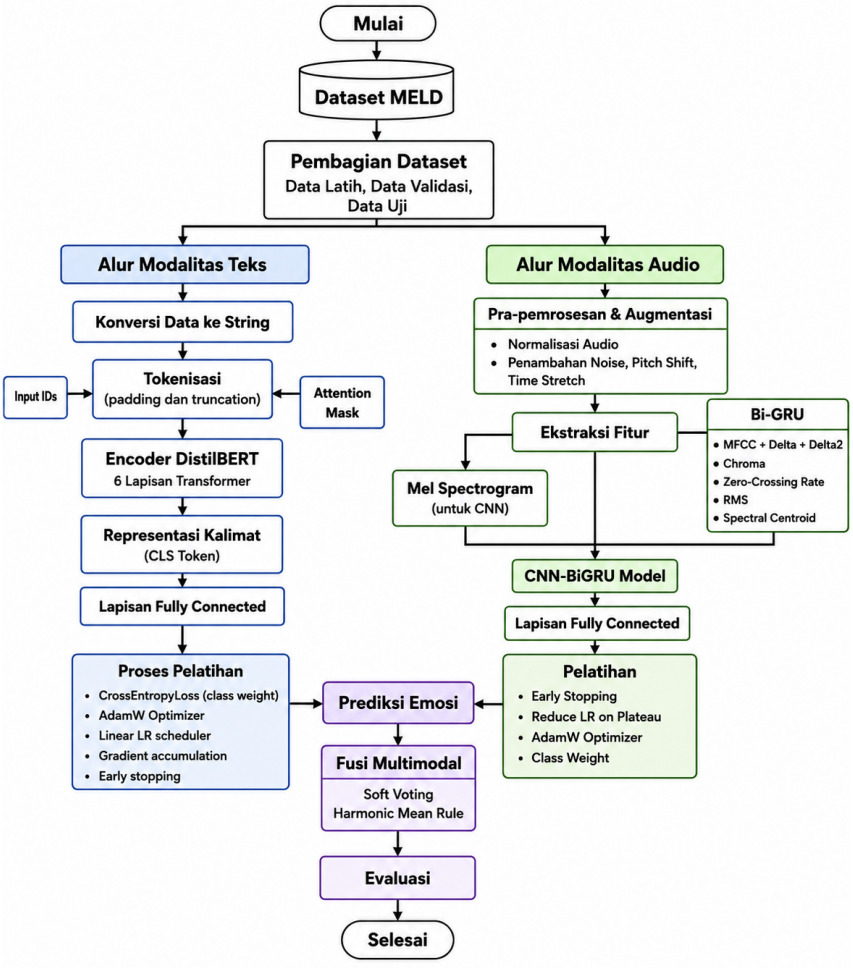
Gambar 1: Distribusi dataset MELD pada data pelatihan, validasi, dan pengujian



Gambar 2: Contoh sampel data MELD

2.2 Alur Penelitian

Penelitian ini mengembangkan klasifikasi emosi multimodal yang menggabungkan informasi tekstual dan akustik dari dataset MELD. Arsitektur yang diusulkan terdiri atas dua cabang utama, yaitu cabang teks dan cabang audio. Diagram alur penelitian ini ditunjukkan pada Gambar 3.



Gambar 3: Alur Penelitian

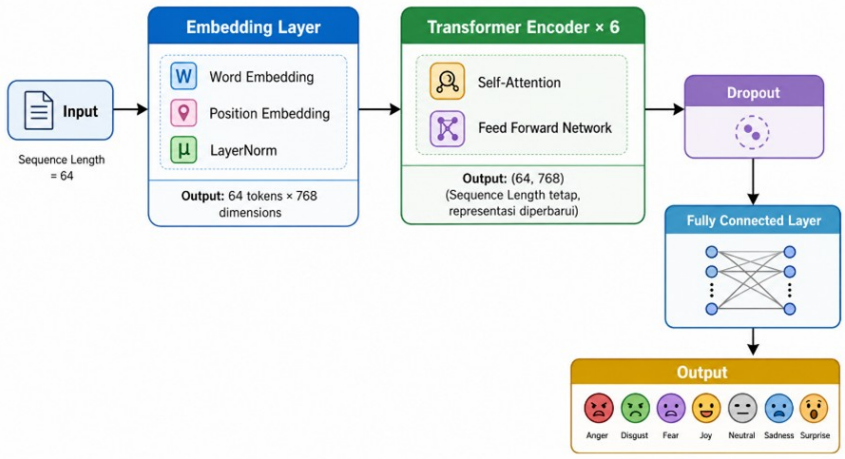
Pada cabang teks, DistilBERT digunakan untuk mengekstraksi representasi kontekstual dari setiap *utterance*. Sementara itu, pada cabang audio dilakukan ekstraksi berbagai fitur akustik yang merepresentasikan karakteristik temporal dan spektral sinyal suara. Fitur audio kemudian diproses menggunakan kombinasi CNN dan BiGRU. Selanjutnya, probabilitas keluaran dari model teks dan audio digabungkan menggunakan pendekatan *ensemble learning* guna menghasilkan prediksi emosi akhir.

2.3 Modalitas Teks

Pada modalitas teks, setiap *utterance* dari dataset MELD digunakan secara langsung tanpa pra-pemrosesan tambahan. Pendekatan ini dipilih karena tokenizer DistilBERT tipe *uncased* telah melakukan normalisasi huruf selama proses tokenisasi. Setiap *utterance* diubah menjadi representasi numerik berupa *input IDs* dan *attention masks*. Proses tokenisasi dilakukan dengan *padding* dan *truncation* untuk menyeragamkan panjang urutan token pada setiap sampel.

2.3.1 Arsitektur DistilBERT

Penelitian ini menggunakan DistilBERT sebagai model klasifikasi pada modalitas teks. DistilBERT merupakan versi ringkas dari BERT yang dikembangkan melalui proses *knowledge distillation*, sehingga memiliki jumlah parameter yang lebih sedikit dan waktu komputasi yang lebih cepat.



Gambar 4: Arsitektur DistilBERT untuk Modalitas Teks

Gambar 4 menunjukkan arsitektur model yang digunakan pada modalitas teks. Setiap *utterance* direpresentasikan menggunakan DistilBERT untuk menghasilkan representasi yang mampu menangkap makna kalimat secara menyeluruh. Proses transformasi teks menjadi representasi fitur dapat dinyatakan sebagai:

$$\mathbf{h} = \text{DistilBERT}(\mathbf{x}) \tag{1}$$

Urutan token masukan dinyatakan sebagai $\mathbf{x}$ dan $\mathbf{h}$ merupakan representasi laten yang dihasilkan oleh lapisan encoder DistilBERT. Representasi tersebut kemudian diteruskan ke lapisan klasifikasi untuk menghasilkan skor bagi setiap kelas emosi sebagai berikut:

$$\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b} \tag{2}$$

dengan $\mathbf{W}$ dan $\mathbf{b}$ masing-masing merupakan bobot dan bias pada lapisan klasifikasi. Selanjutnya, probabilitas setiap kelas emosi dihitung menggunakan fungsi *softmax*:

$$P(y = i|\mathbf{x}) = \frac{\exp(z_i)}{\sum_{j=1}^{7} \exp(z_j)} \tag{3}$$

$P(y = i|\mathbf{x})$ menyatakan probabilitas bahwa suatu *utterance* termasuk ke dalam kelas emosi ke-i dari tujuh kelas emosi pada dataset MELD. Kelas emosi dengan probabilitas tertinggi dipilih sebagai hasil prediksi akhir. Output dari model teks ini kemudian digunakan pada tahap *ensemble* bersama hasil prediksi dari modalitas audio untuk menghasilkan keputusan klasifikasi multimodal.

2.4 Modalitas Audio

Pada modalitas audio, setiap *utterance* diproses menggunakan Librosa untuk mengekstraksi fitur akustik. Penelitian ini memanfaatkan dua jenis representasi audio, yaitu fitur *sequential* dan fitur *spectral*. Fitur *sequential* terdiri atas MFCC, Chroma, Spectral Centroid, Zero Crossing Rate (ZCR), dan Root Mean Square Energy (RMSE), sedangkan fitur *spectral* direpresentasikan dalam bentuk *Mel-Spectrogram*.

Untuk mengatasi ketidakseimbangan distribusi kelas, dilakukan kombinasi *undersampling* pada kelas mayoritas dan *audio augmentation* pada kelas minoritas. Teknik augmentasi yang digunakan meliputi penambahan *noise*, *pitch shifting*, dan *time stretching*. Hasil ekstraksi fitur kemudian digunakan sebagai masukan pada model klasifikasi audio yang terdiri atas cabang CNN untuk memproses *Mel-Spectrogram* dan cabang BiGRU untuk memproses fitur *sequential*.

Pada cabang CNN, proses ekstraksi fitur dilakukan melalui operasi konvolusi yang dinyatakan sebagai:

$$y_j = f\left(\sum_i x_i \times w_{ij} + b_j\right) \tag{4}$$

Pada Persamaan 4, x_i merupakan fitur masukan, w_{ij} adalah kernel konvolusi, b_j adalah bias, dan $f(\cdot)$ merupakan fungsi aktivasi. Operasi ini memungkinkan model mengekstraksi pola lokal pada representasi *Mel-Spectrogram*. Sementara itu, cabang BiGRU digunakan untuk menangkap dependensi temporal dari fitur *sequential*. Hidden state pada GRU dihitung sebagaimana Persamaan 5:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{5}$$

Hidden state pada waktu ke-t dituliskan sebagai h_t, sedangkan z_t adalah *update gate*, dan $\tilde{h}_t$ merupakan kandidat hidden state. Representasi yang dihasilkan oleh kedua cabang kemudian digabungkan dan digunakan untuk menghasilkan prediksi emosi pada modalitas audio.

2.5 Fusi Multimodal

Fusi multimodal dilakukan dengan menggabungkan probabilitas keluaran dari model teks dan model audio. Pada penelitian ini digunakan dua pendekatan *ensemble learning*, yaitu *soft voting* dan *harmonic mean* untuk menghasilkan prediksi emosi akhir berdasarkan informasi dari kedua

modalitas. Metode *soft voting* menghitung rata-rata probabilitas keluaran dari modalitas teks dan audio untuk setiap kelas emosi. Formulasi *soft voting* dinyatakan pada Persamaan 6, dengan P_t dan P_a masing-masing menyatakan probabilitas keluaran dari model teks dan model audio.

$$P_{sv} = \frac{P_t + P_a}{2} \tag{6}$$

Selain *soft voting*, penelitian ini juga menerapkan *harmonic mean* untuk menggabungkan probabilitas dari kedua modalitas. Metode ini memberikan bobot yang lebih rendah ketika salah satu modalitas menghasilkan probabilitas yang kecil. Formulasi *harmonic mean* ditunjukkan pada Persamaan 7. Nilai probabilitas hasil fusi digunakan untuk menentukan kelas emosi akhir berdasarkan probabilitas tertinggi pada masing-masing sampel.

$$H = \frac{2 \times P_t \times P_a}{P_t + P_a} \tag{7}$$

2.6 Metrik Evaluasi

Kinerja model dievaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, *Macro F1-score*, dan *Weighted F1-score*. Nilai *accuracy*, *precision*, *recall*, dan *F1-score* dihitung menggunakan Persamaan (8)–(11). Selanjutnya, *Macro F1-score* dihitung sebagai rata-rata nilai *F1-score* dari seluruh kelas, sedangkan *Weighted F1-score* dihitung berdasarkan rata-rata *F1-score* yang mempertimbangkan jumlah sampel pada setiap kelas.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Presisi} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$F1\text{-Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \tag{11}$$

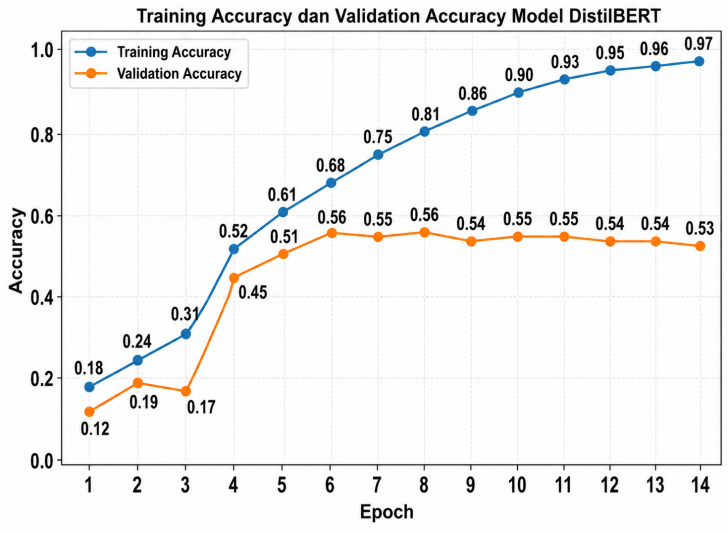
dengan TP, TN, FP, dan FN masing-masing menyatakan *True Positive*, *True Negative*, *False Positive*, dan *False Negative*. Seluruh metrik tersebut digunakan untuk mengevaluasi performa model dari berbagai aspek.

3 Hasil dan Pembahasan

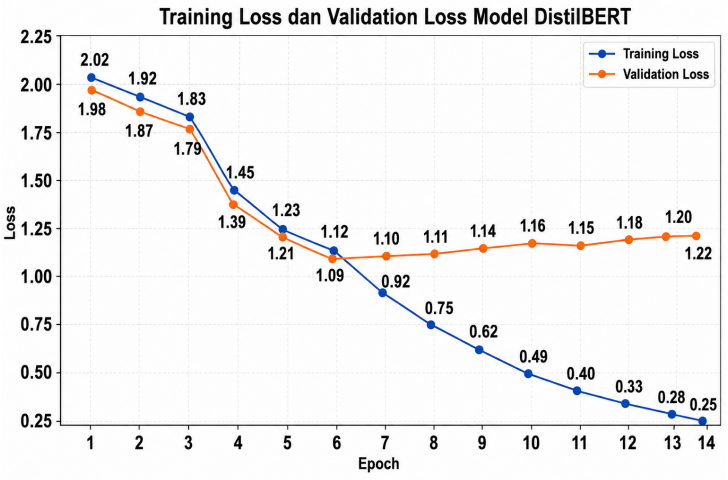
3.1 Kinerja Modalitas Teks

Model DistilBERT menghasilkan akurasi sebesar 63% pada data pengujian, dengan *Macro F1-score* sebesar 45,3% dan *Weighted F1-score* sebesar 62%. Hasil tersebut menunjukkan bahwa modalitas teks memberikan performa yang lebih baik dibandingkan modalitas audio dalam klasifikasi emosi pada dataset MELD. Selain itu, nilai *Macro F1-score* yang lebih rendah dibandingkan *Weighted F1-score* mengindikasikan bahwa performa model belum sepenuhnya merata pada seluruh kelas emosi. Untuk memberikan gambaran mengenai proses pembelajaran model selama pelatihan,

Gambar 5 menyajikan kurva *training accuracy*, *validation accuracy*, *training loss*, dan *validation loss*.



(a) Grafik Akurasi Pelatihan dan Validasi



(b) Grafik Loss Pelatihan dan Validasi

Gambar 5: Kurva Akurasi dan Loss Pelatihan serta Validasi Model DistilBERT

Berdasarkan Gambar 5, nilai *training accuracy* dan *validation accuracy* menunjukkan peningkatan yang konsisten selama proses pelatihan, sedangkan *training loss* dan *validation loss* mengalami penurunan secara bertahap. Selisih antara kurva pelatihan dan validasi relatif kecil sehingga tidak terlihat indikasi *overfitting* yang signifikan. Kondisi tersebut menunjukkan bahwa model mampu mempelajari pola pada data pelatihan sekaligus mempertahankan kemampuan generalisasi yang baik pada data validasi. Hal ini menunjukkan bahwa representasi linguistik yang dipelajari oleh DistilBERT cukup efektif untuk membedakan kategori emosi pada dataset MELD.

Tabel 1: Hasil Evaluasi Klasifikasi Modalitas Teks Menggunakan DistilBERT

Emosi	**Presisi**	**Recall**	**F1-Score**	**Jumlah Data**
Anger	0.50	0.34	0.41	345
Disgust	0.26	0.26	0.26	68
Fear	0.29	0.16	0.21	50
Joy	0.55	0.63	0.59	402
Neutral	0.37	0.31	0.34	1256
Sadness	0.53	0.60	0.57	208
Surprise	0.77	0.81	0.79	281
Macro Avg	0.47	0.44	0.45	2610
Weighted Avg	0.62	0.63	0.62	2610
Akurasi			0.63	2610

Berdasarkan Tabel 1, kelas *surprise* memperoleh performa terbaik dengan nilai F1-score sebesar 0,79, diikuti oleh kelas *joy* dan *sadness* dengan nilai *F1-score* masing-masing sebesar 0,59 dan 0,57. Hasil tersebut menunjukkan bahwa emosi yang memiliki karakteristik linguistik lebih jelas cenderung lebih mudah dikenali oleh model. Sebaliknya, kelas *fear* dan *disgust* masih memperoleh nilai *F1-score* yang relatif rendah, masing-masing sebesar 0,21 dan 0,26. Selain itu, nilai *Macro F1-score* sebesar 45,3% yang lebih rendah dibandingkan *Weighted F1-score* sebesar 62% menunjukkan bahwa performa model belum merata pada seluruh kelas emosi.

3.2 Kinerja Modalitas Audio

Model audio berbasis CNN-BiGRU menghasilkan akurasi sebesar 46% pada data pengujian, dengan *Macro F1-score* sebesar 14,4% dan *Weighted F1-score* sebesar 35%. Hasil tersebut menunjukkan bahwa model masih mengalami kesulitan dalam mengenali seluruh kelas emosi secara merata. Selain itu, performa modalitas audio masih berada di bawah modalitas teks, yang mengindikasikan bahwa informasi akustik pada dataset MELD lebih sulit dimanfaatkan untuk membedakan kategori emosi.

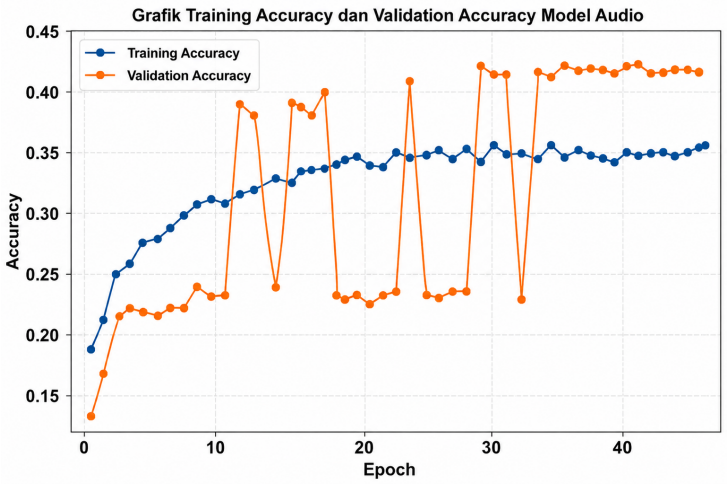
Tabel 2: Hasil Evaluasi Klasifikasi Modalitas Audio

Emosi	**Presisi**	**Recall**	**F1-Score**	**Jumlah Data**
Anger	0.33	0.11	0.16	345
Disgust	0.06	0.09	0.07	68
Fear	0.00	0.00	0.00	50
Joy	0.29	0.03	0.05	402
Neutral	0.50	0.90	0.64	1256
Sadness	0.09	0.01	0.02	208
Surprise	0.22	0.04	0.07	281
Macro Avg	0.21	0.17	0.14	2610
Weighted Avg	0.36	0.46	0.35	2610
Akurasi			0.46	2610

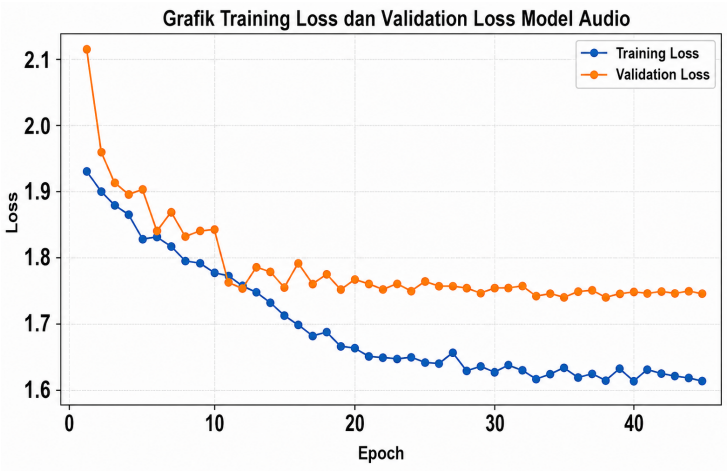
Berdasarkan Tabel 2, model menunjukkan performa yang bervariasi pada setiap kelas emosi. Kelas *neutral* memperoleh performa terbaik dengan nilai *recall* sebesar 0,90 dan *F1-score* sebesar

0,64, sedangkan kelas *fear* tidak berhasil dikenali sama sekali dengan nilai *precision*, *recall*, dan *F1-score* sebesar 0,00. Selain itu, kelas *disgust*, *joy*, *sadness*, dan *surprise* juga memperoleh nilai *F1-score* yang relatif rendah. Nilai *Macro F1-score* sebesar 14,4% yang jauh lebih rendah dibandingkan *Weighted F1-score* sebesar 35% menunjukkan bahwa performa model belum merata pada seluruh kelas emosi. Kondisi ini mengindikasikan bahwa model cenderung lebih baik dalam mengenali kelas dengan jumlah sampel yang besar dibandingkan kelas minoritas.

Untuk memberikan gambaran mengenai proses pembelajaran model selama pelatihan, Gambar 6 menyajikan kurva *training accuracy*, *validation accuracy*, *training loss*, dan *validation loss*.



(a) Grafik Akurasi Pelatihan dan Validasi



(b) Grafik Loss Pelatihan dan Validasi

Gambar 6: Kurva Akurasi dan Loss Pelatihan serta Validasi Model CNN-BiGRU

Berdasarkan Gambar 6, nilai *training accuracy* menunjukkan peningkatan secara bertahap selama proses pelatihan, sedangkan *validation accuracy* masih berfluktuasi, terutama pada rentang epoch 10–20 dan 20–30. Fluktuasi tersebut menunjukkan bahwa model belum mampu menghasilkan performa yang konsisten pada data validasi. Kondisi ini juga didukung oleh kurva

validation loss yang cenderung stagnan setelah mengalami penurunan pada awal pelatihan, sehingga peningkatan kemampuan model pada data pelatihan tidak diikuti oleh peningkatan performa pada data validasi. Temuan tersebut mengindikasikan bahwa kemampuan generalisasi model masih terbatas, yang diduga dipengaruhi oleh variasi karakteristik suara antar pembicara serta ketidakseimbangan distribusi kelas pada dataset MELD.

3.3 Perbandingan Kinerja Model Unimodal dan Multimodal

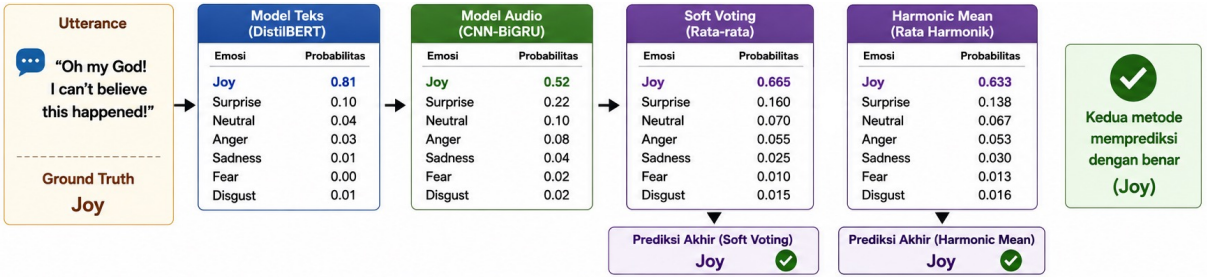
Perbandingan kinerja seluruh model ditunjukkan pada Tabel 3. Hasil pengujian menunjukkan bahwa model audio berbasis CNN-BiGRU memperoleh performa terendah dengan akurasi sebesar 46%, *Macro F1-score* sebesar 14,4%, dan *Weighted F1-score* sebesar 35%. Sebaliknya, model teks berbasis DistilBERT menghasilkan akurasi sebesar 63%, *Macro F1-score* sebesar 45,3%, dan *Weighted F1-score* sebesar 62%, yang menunjukkan bahwa informasi tekstual memiliki kontribusi lebih besar dalam proses klasifikasi emosi pada dataset MELD.

Tabel 3: Perbandingan Kinerja Model Unimodal dan Multimodal

Model	Macro F1-Score	Weighted F1-Score	Accuracy
Audio (CNN-BiGRU)	14.4%	35%	46%
Text (DistilBERT)	45.3%	62%	63%
Fusion (Soft Voting)	45.3%	62%	64%
Fusion (Harmonic Mean)	44.6%	62%	65%

Penggabungan modalitas teks dan audio melalui pendekatan *ensemble* menghasilkan peningkatan nilai *accuracy* dibandingkan model teks tunggal. Metode *soft voting* memperoleh akurasi sebesar 64% dengan *Macro F1-score* sebesar 45,3% dan *Weighted F1-score* sebesar 62%. Sedangkan metode *harmonic mean* menghasilkan akurasi tertinggi sebesar 65% dengan *Macro F1-score* sebesar 44,6% dan *Weighted F1-score* sebesar 62%.

Meskipun pendekatan multimodal meningkatkan akurasi, nilai *Macro F1-score* dan *Weighted F1-score* tidak mengalami peningkatan dibandingkan model teks tunggal. Hal ini menunjukkan bahwa penggabungan modalitas audio belum mampu meningkatkan kemampuan model dalam mengenali seluruh kelas emosi secara lebih merata. Namun demikian, metode *harmonic mean* tetap menghasilkan akurasi tertinggi sebesar 65%, lebih tinggi dibandingkan metode *soft voting* maupun model DistilBERT. Oleh karena itu, metode *harmonic mean* dipilih sebagai model terbaik pada penelitian ini karena memberikan performa keseluruhan terbaik berdasarkan metrik akurasi. Untuk memberikan gambaran mengenai proses pengambilan keputusan pada model *ensemble*, Gambar 7 menyajikan contoh hasil prediksi.



Gambar 7: Contoh hasil prediksi menggunakan metode *soft voting* dan *harmonic mean*.

Berdasarkan Gambar 7, model teks dan model audio menghasilkan distribusi probabilitas yang kemudian digabungkan menggunakan metode *soft voting* dan *harmonic mean*. Pada contoh tersebut, kedua metode *ensemble* menghasilkan prediksi yang sesuai dengan *ground truth*. Namun, pada beberapa sampel lain prediksi masih dapat berbeda dari *ground truth* apabila salah satu modalitas memberikan probabilitas yang lebih dominan pada kelas yang kurang tepat.

3.4 Perbandingan dengan Penelitian Sebelumnya

Perbandingan metode yang diusulkan dengan beberapa penelitian terdahulu pada dataset MELD ditunjukkan pada Tabel 4–6. Penelitian-penelitian sebelumnya menggunakan berbagai pendekatan untuk klasifikasi emosi, baik pada modalitas audio, teks, maupun multimodal. Pada penelitian ini, perbandingan dilakukan menggunakan tiga metrik evaluasi, yaitu *Accuracy*, *Weighted F1-score*, dan *Macro F1-score*, sehingga dapat memberikan gambaran yang lebih komprehensif mengenai performa model dalam mengklasifikasikan emosi.

Perlu diperhatikan bahwa hasil perbandingan tidak sepenuhnya bersifat setara karena setiap penelitian menggunakan konfigurasi eksperimen yang berbeda, seperti pembagian data, strategi prapemrosesan, arsitektur model, jumlah epoch, serta teknik optimasi yang digunakan. Oleh karena itu, perbandingan ini bertujuan untuk memberikan gambaran posisi performa metode yang diusulkan terhadap penelitian terdahulu pada dataset yang sama, bukan sebagai perbandingan langsung.

Tabel 4: Perbandingan dengan Penelitian Sebelumnya Menggunakan Metrik Akurasi

Peneliti	**Modalitas**	**Metode**	**Akurasi**
He (2026)	A	IGEM	40%
	T	IGEM	46%
	T+A	IGEM	48%
Wang and Beigi (2025)	A	MAMBA Fusion	41%
	T	MPNet-v2	62%
	T+A	MPNet-v2 + MAMBA Fusion	65%
Adel et al. (2023)	A	MM-EMOR	45%
	T	MM-EMOR	62%
	T+A	MM-EMOR	63%
Aguilera et al. (2023)	A	Multimodal Deep Learning	43%
	T	Multimodal Deep Learning	57%
Luo et al. (2023)	A	CM-RoBERTa	42%
	T	CM-RoBERTa	64%
	T+A	CM-RoBERTa	64%
Jiao et al. (2020)	T	UniF-BiAGRU	60%
Penelitian Ini	A	CNN–BiGRU	46%
	T	DistilBERT	63%
	T+A	DistilBERT + CNN–BiGRU + Soft Voting	64%
	T+A	DistilBERT + CNN–BiGRU + Harmonic Mean	65%

Tabel 5: Perbandingan dengan Penelitian Sebelumnya Menggunakan Metrik WF1

Peneliti	**Modalitas**	**Metode**	**Weighted F1-Score**
Hu et al. (2021)	A	MMGCN	42%
	T	MMGCN	55%
	T+A	MMGCN	58%
Lian et al. (2021)	A	CTNet	39%
	T	CTNet	59%
	T+A	CTNet	61%
Ghosal et al. (2019)	A	DialogueGCN	41%
	T	DialogueGCN	58%
	T+A	DialogueGCN	60%
Penelitian Ini	A	CNN–BiGRU	35%
	T	DistilBERT	62%
	T+A	DistilBERT + CNN–BiGRU + Soft Voting	62%
	T+A	DistilBERT + CNN–BiGRU + Harmonic Mean	62%

Tabel 6: Perbandingan dengan Penelitian Sebelumnya Menggunakan Metrik Macro F1-Score

Peneliti	**Modalitas**	**Metode**	**Macro F1-Score**
Ong et al. (2024)	T	RoBERTa + Contrastive Learning + Emotional Intensity	49.68%
Aguilera et al. (2023)	A	Multimodal Deep Learning Architecture	34.6%
Jiao et al. (2020)	T	UniF-BiAGRU	38.6%
Penelitian Ini	A	CNN–BiGRU	14.4%
	T	DistilBERT	45.3%
	T+A	DistilBERT + CNN–BiGRU + Soft Voting	45.3%
	T+A	DistilBERT + CNN–BiGRU + Harmonic Mean	44.6%

4 Kesimpulan

Penelitian ini mengevaluasi pengenalan emosi pada dataset MELD menggunakan modalitas teks, audio, dan multimodal. Hasil pengujian menunjukkan bahwa model berbasis teks menggunakan DistilBERT memberikan performa yang lebih baik dibandingkan model berbasis audio CNN–BiGRU. Model teks memperoleh akurasi sebesar 63%, *Macro F1-score* sebesar 45,3%, dan *Weighted F1-score* sebesar 62%, sedangkan model audio memperoleh akurasi sebesar 46%, *Macro F1-score* sebesar 14,4%, dan *Weighted F1-score* sebesar 35%. Hasil tersebut menunjukkan bahwa informasi tekstual memiliki kontribusi yang lebih besar dibandingkan informasi akustik dalam membedakan kategori emosi pada dataset MELD.

Penggabungan modalitas teks dan audio melalui metode *ensemble* meningkatkan akurasi

menjadi 65% menggunakan metode *harmonic mean*, sedangkan metode *soft voting* menghasilkan akurasi sebesar 64%. Meskipun demikian, kedua metode tersebut belum mampu meningkatkan *Macro F1-score* maupun *Weighted F1-score* dibandingkan model DistilBERT. Dengan demikian, metode *harmonic mean* dipilih sebagai pendekatan terbaik pada penelitian ini karena menghasilkan akurasi tertinggi dalam klasifikasi emosi multimodal. Hasil penelitian ini menunjukkan bahwa pendekatan *ensemble* sederhana mampu meningkatkan performa klasifikasi multimodal tanpa memerlukan arsitektur fusi yang kompleks, meskipun peningkatan tersebut masih terbatas pada metrik akurasi. Oleh karena itu, penelitian selanjutnya dapat mengeksplorasi strategi penanganan ketidakseimbangan kelas maupun metode fusi multimodal yang lebih efektif untuk meningkatkan performa pada seluruh kategori emosi.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Program Studi Sains Data Universitas Negeri Surabaya serta seluruh pihak yang telah memberikan dukungan dan masukan selama pelaksanaan penelitian ini.

References

Adel, O., Fathalla, K. M., and Abo ElFarag, A. (2023). Mm-emor: Multi-modal emotion recognition of social media using concatenated deep learning networks. *Big Data and Cognitive Computing*, 7(4):164.

Aguilera, A., Mellado, D., and Rojas, F. (2023). An assessment of in-the-wild datasets for multimodal emotion recognition. *Sensors*, 23(11):5184.

Dadeh, T., Alfarisi, M. R., Hutasoit, G., and Rasool, S. (2025). The impact of school climate on emotion regulation in indonesian students: Evidence from pisa 2022. *Journal An-Nafs*, 10(1):27–45.

Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP*, pages 154–164.

Harahap, Y., Sari, M., Nurjannah, V., et al. (2024). Inovasi emokids: Alat pendeteksi emosi pada anak berbasis image processing. *BEST Journal*, 7(2):1376–1382.

He, J. (2026). Interactive graph emotion recognition based on multi-modal data enhancement. *Scientific Reports*, 16(1).

Higuchi, M., Nakamura, M., Shinohara, S., et al. (2020). Effectiveness of a voice-based mental health evaluation system for mobile devices. *JMIR Formative Research*, 4(7):e16455.

Hu, J., Liu, Y., Zhao, J., and Jin, Q. (2021). Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *ACL*, pages 5666–5675.

Jayaswal, R., Ansari, M. A., Dixit, M., Singh, D. K., and Ahmad, S. (2025). Advances in facial expression recognition technologies for emotion analysis. *Discover Computing*, 28(1).

Jiao, W., Lyu, M., and King, I. (2020). Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8002–8009.

Jordan, E., Terrisse, R., Lucarini, V., et al. (2025). Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR Mental Health*, 12:e74260.

Larrouy-Maestri, P., Poeppel, D., and Pell, M. D. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*, 20(4):623–638.

Lian, Z., Liu, B., and Tao, J. (2021). Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM TASLP*, 29:985–1000.

Luo, J., Phan, H., and Reiss, J. (2023). Cross-modal fusion techniques for utterance-level emotion recognition. In *ICASSP*, pages 1–5.

Martinez, A. M. (2017). Visual perception of facial expressions of emotion. *Current Opinion in Psychology*, 17:27–33.

Ong, D., Sun, S., Su, J., and Chen, B. (2024). Mitigating linguistic artifacts in emotion recognition for conversations: From tv scripts to daily conversations. In *Proceedings of the Language Resources and Evaluation Conference (LREC-COLING 2024)*, pages 11319–11324.

Poria, S., Hazarika, D., Majumder, N., et al. (2019). Meld dataset for emotion recognition in conversations. In *ACL*.

Tang, V. (2025). Usc secures hat trick at interspeech 2025. *USC Viterbi News*.

Tang, Y. and He, W. (2023). Relationship between emotional intelligence and learning motivation among college students during the covid-19 pandemic. *Frontiers in Psychology*, 14.

Wang, Z. and Beigi, H. (2025). Multimodal emotion recognition with mamba fusion. *arXiv*.