STARNAS 2025
Seminar Teknologi dan Sains Nasional

Analisis Statistik Dataset Emosi Bahasa Indonesia untuk Pengembangan Model Klasifikasi Emosi

[1]Yoppy Sazaki, [2]Mgs. Afriyan Firdaus, [3]Junia Kurniati
[1,2,3] Fakultas Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia
Corresponding Author: yoppysazaki@gmail.com

Abstract
This study presents a descriptive statistical analysis of an Indonesian emotion dataset as an essential foundation for developing emotion classification models in Natural Language Processing (NLP). A representative and balanced dataset is crucial to ensuring stable and unbiased model predictions. However, many existing Indonesian emotion datasets exhibit class imbalance, limiting the model's ability to learn minority emotion patterns. The dataset analyzed in this study consists of 7,368 text samples annotated into six emotion categories: angry, disgusted, fearful, happy, neutral, and sad. In addition to emotion distribution, this research also examines text-length characteristics to better understand the linguistic structure of the dataset, as text length can influence the performance of modern NLP models. The analysis reveals that sad, happy, and neutral are the dominant categories, while angry and disgusted appear in smaller proportions. Such imbalance may lead to biased predictions toward majority classes during model training. These findings provide an initial understanding of the statistical characteristics of Indonesian emotion data and highlight the importance of applying data-balancing strategies in future research to improve the performance and fairness of emotion classification models.
Keywords: *statistical analysis; emotion dataset; data imbalance; emotion classification; Indonesian language; natural language processing*

Abstrak
Penelitian ini menyajikan analisis statistik deskriptif terhadap dataset emosi Bahasa Indonesia sebagai dasar awal dalam pengembangan model klasifikasi emosi berbasis pemrosesan bahasa alami. Dataset yang representatif dan seimbang sangat penting untuk menghasilkan prediksi model yang stabil dan tidak bias. Namun, banyak dataset emosi yang tersedia menunjukkan ketidakseimbangan kelas yang dapat mengurangi kemampuan model untuk mempelajari pola emosi minoritas. Dataset yang dianalisis dalam penelitian ini terdiri atas 7.368 sampel teks yang diberi label ke dalam enam kategori emosi, yaitu angry, disgusted, fearful, happy, neutral, dan sad. Selain distribusi emosi, penelitian ini juga menganalisis karakteristik panjang teks untuk memahami struktur linguistik dataset, karena faktor ini berpengaruh terhadap kinerja model NLP modern. Hasil analisis menunjukkan bahwa kategori sad, happy, dan neutral mendominasi dataset, sedangkan angry dan disgusted memiliki proporsi yang lebih kecil. Ketidakseimbangan ini berpotensi menyebabkan bias prediksi terhadap kelas mayoritas pada saat pelatihan model. Temuan ini memberikan gambaran awal mengenai karakteristik statistik dataset emosi Bahasa Indonesia serta menegaskan pentingnya penerapan strategi penyeimbangan data dalam penelitian lanjutan untuk meningkatkan akurasi dan keadilan model klasifikasi emosi.
Kata-Kata Kunci: *analisis statistik; dataset emosi; ketidakseimbangan data; klasifikasi emosi; bahasa Indonesia; pemrosesan bahasa alami*.

1. PENDAHULUAN

Analisis emosi pada teks berbahasa Indonesia semakin penting seiring meningkatnya penggunaan media sosial, layanan *chatbot*, serta aplikasi berbasis kecerdasan buatan yang membutuhkan pemahaman terhadap kondisi emosional pengguna. Studi global menunjukkan bahwa pemodelan emosi pada teks memiliki peran penting dalam peningkatan interaksi manusia–mesin dan deteksi kondisi psikologis (Zhang et al., 2022; Üveges & Ring, 2025). Untuk menghasilkan model klasifikasi emosi yang akurat dan tidak bias, diperlukan dataset yang representatif dan seimbang sehingga mampu mencerminkan variasi emosi secara proporsional (Yosef et al., 2021; Kesanam et al., 2025). Namun, berbagai dataset emosi berbahasa Indonesia masih menunjukkan ketidakseimbangan distribusi antar kategori, yang dapat mengakibatkan bias prediksi terhadap kelas mayoritas (Wang & Culotta, 2020; Fernandes et al., 2023).

Sebagian besar penelitian sebelumnya berfokus pada pengembangan model klasifikasi emosi dengan memanfaatkan pendekatan berbasis *deep learning* dan Transformer seperti BERT, RoBERTa, dan DistilBERT (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2020). Di Indonesia, penelitian terkait NLP mulai banyak memanfaatkan model pralatih seperti IndoBERT dan RoBERTa-Ind untuk berbagai tugas analisis emosi (Koto et al., 2020; Wilie et al., 2020). Namun, penelitian-penelitian tersebut tidak melakukan analisis statistik komprehensif terhadap dataset, sehingga potensi ketidakseimbangan data dan implikasinya terhadap kinerja model tidak terlihat (Al-Azani et al., 2025). Padahal, riset terkini menunjukkan bahwa *class imbalance* merupakan faktor kunci yang menentukan keberhasilan model klasifikasi emosi (Buda et al., 2018; Radliński et al., 2025).

Kondisi ini memunculkan research gap yang penting: minimnya penelitian yang secara khusus menganalisis karakteristik statistik dataset emosi Bahasa Indonesia sebelum digunakan dalam pelatihan model klasifikasi emosi. Untuk itu, penelitian ini mengajukan beberapa pertanyaan: (1) bagaimana distribusi kategori emosi dalam dataset Bahasa Indonesia? (2) Seberapa besar tingkat ketidakseimbangan antar kelas? (3) Bagaimana implikasinya terhadap proses pengembangan model klasifikasi emosi? Pertanyaan ini penting untuk memastikan model NLP tidak hanya akurat, tetapi juga adil terhadap seluruh kelas emosi.

Penelitian ini bertujuan untuk melakukan analisis statistik deskriptif terhadap dataset emosi Bahasa Indonesia guna mengidentifikasi proporsi distribusi kelas emosi serta mengevaluasi tingkat ketidakseimbangan data. Temuan ini menjadi dasar penting bagi penelitian NLP lanjutan, terutama dalam merancang strategi penyeimbangan data seperti *oversampling*, *undersampling*, *class weighting*, atau penggunaan fungsi *loss* khusus seperti Focal Loss (Lin et al., 2017; Ye et al., 2022).

Novelty penelitian ini terletak pada analisis statistik menyeluruh terhadap dataset emosi Bahasa Indonesia menggunakan pendekatan evaluatif yang didukung referensi mutakhir, termasuk kontribusi penelitian terbaru tahun 2025 mengenai ketidakseimbangan data dan klasifikasi emosi (Kesanam et al.,

2025; Radliński et al., 2025; Üveges & Ring, 2025). Pendekatan ini berbeda dari penelitian sebelumnya yang langsung membangun model tanpa mengevaluasi kualitas dataset terlebih dahulu.

Penelitian ini memberikan manfaat teoretis berupa pemahaman struktur dataset emosi Bahasa Indonesia, serta manfaat praktis berupa rekomendasi strategi penanganan ketidakseimbangan data untuk meningkatkan akurasi dan keadilan model emosi. Korelasi antara *research problem* (ketidakseimbangan data), *research questions* (karakteristik distribusi emosi), dan *research objectives* (analisis statistik dan rekomendasi balancing) menunjukkan bahwa penelitian ini disusun secara sistematis untuk mendukung pengembangan model klasifikasi emosi yang lebih robust.

2. KERANGKA TEORI

Analisis emosi pada teks merupakan cabang penting dalam pemrosesan bahasa alami (Natural Language Processing/NLP) yang bertujuan mengidentifikasi ekspresi emosional dalam suatu teks. Emosi pada teks umumnya direpresentasikan dalam bentuk kategori tertentu seperti *angry*, *happy*, *sad*, atau *neutral*. Penelitian sebelumnya menyatakan bahwa analisis emosi memiliki peran penting dalam aplikasi seperti deteksi kesehatan mental, analisis opini publik, dan interaksi manusia–mesin (Zhang et al., 2022; Akhtar & Mittal, 2023). Pada tataran NLP modern, tugas klasifikasi emosi secara umum memerlukan dataset yang memadai, terstruktur, dan memiliki distribusi kelas yang representatif (Üveges & Ring, 2025).

Selain dibedakan ke dalam kategori umum seperti *angry*, *happy*, *sad*, atau *neutral*, penelitian psikologi modern menjelaskan bahwa emosi memiliki struktur dasar yang telah banyak digunakan sebagai acuan dalam penelitian NLP. Salah satu model yang paling berpengaruh adalah *basic emotions* yang dikemukakan oleh Paul Ekman, yang mengidentifikasi enam emosi utama yang dianggap universal dan muncul secara konsisten pada berbagai budaya. Model ini sering menjadi dasar dalam pengembangan skema anotasi atau kategori emosi dalam dataset, termasuk yang digunakan dalam penelitian NLP berbahasa Indonesia (Ekman, 1971; Chen & Li, 2024). Enam emosi tersebut ditunjukkan pada Tabel 1.

Tabel 1. Enam Emosi Dasar Menurut Ekman (1971)

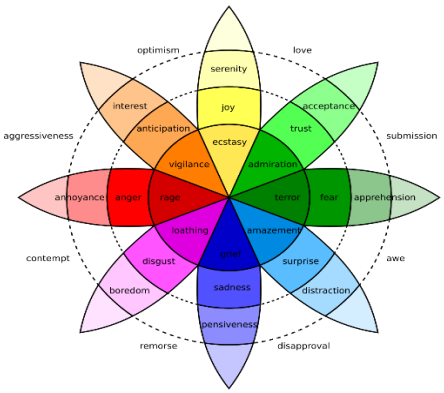
Emosi	Deskripsi Singkat
Angry	Respons terhadap ancaman, frustrasi, atau ketidakadilan
Disgusted	Respons terhadap hal yang menjijikkan atau tidak menyenangkan
Fearful	Respons terhadap bahaya atau ancaman
Happy	Kondisi emosional positif, rasa senang
Sad	Emosi negatif akibat kehilangan atau kekecewaan
Neutral	Tidak menunjukkan emosi dominan

Dataset merupakan komponen fundamental dalam pembangunan model NLP. Kualitas dataset akan menentukan kemampuan model dalam mengenali pola linguistik yang berguna untuk melakukan prediksi pada data baru. Menurut Kesanam et al. (2025), distribusi dataset yang tidak seimbang atau *class imbalance* merupakan salah satu tantangan utama dalam pengembangan model klasifikasi emosi. *Class imbalance* terjadi apabila satu atau beberapa kelas memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas lainnya. Kondisi ini menyebabkan model cenderung belajar lebih banyak dari kelas mayoritas sehingga prediksi menjadi bias dan kurang akurat untuk kelas minoritas (Buda et al., 2018; Wang & Culotta, 2020).

Fenomena ketidakseimbangan data sangat umum ditemukan pada dataset emosi, baik dalam bahasa Indonesia maupun bahasa lain. Hal ini disebabkan oleh kecenderungan alami pengguna untuk mengekspresikan emosi tertentu lebih sering daripada emosi lainnya. Penelitian Radliński et al. (2025) menunjukkan bahwa ketidakseimbangan data berpengaruh signifikan terhadap penurunan performa model Transformer pada tugas klasifikasi emosi. Selain itu, studi Musaev et al. (2024) menegaskan bahwa model berbasis *deep neural networks* sangat sensitif terhadap distribusi kelas yang tidak seimbang dan memerlukan strategi penyeimbangan khusus untuk meningkatkan performa.

Dalam konteks NLP bahasa Indonesia, model pralatih seperti IndoBERT dan RoBERTa-Ind telah digunakan secara luas untuk berbagai tugas analisis teks (Koto et al., 2020; Wilie et al., 2020). Namun sebagian besar penelitian ini langsung fokus pada pelatihan model tanpa melakukan analisis statistik dataset secara mendalam. Padahal pemahaman karakteristik dataset merupakan langkah awal yang penting untuk merancang strategi pelatihan yang lebih baik (Al-Azani et al., 2025). Kajian dataset dapat mencakup analisis distribusi kelas, panjang teks, variasi linguistik, serta hubungan antar kategori emosi (Chen & Li, 2024).

Dalam teori psikologi modern, struktur emosi juga digambarkan menggunakan *Plutchik's Wheel of Emotions*, yang memvisualisasikan hubungan intensitas dan kombinasi emosi (Plutchik, 1980). Model ini membantu memahami keterkaitan semantik antar emosi yang dapat digunakan sebagai acuan dalam pembangunan model NLP dan analisis emosi lebih lanjut. Gambar 1 menunjukkan struktur roda emosi tersebut.



Gambar 1. Plutchik's Wheel of Emotions

Strategi umum untuk mengatasi ketidakseimbangan data meliputi teknik *oversampling*, *undersampling*, *synthetic data generation*, penerapan *class weighting*, serta penggunaan fungsi *loss* khusus seperti Cross-Entropy dan Focal Loss (Lin et al., 2017; Ye et al., 2022). Cross-Entropy Loss merupakan fungsi dasar pada klasifikasi emosi:

$$L = -\sum_{i=1}^{N} y_i \log(p_i) \quad (1)$$

dimana y_i adalah label sebenarnya dan p_i adalah probabilitas prediksi model. Namun fungsi ini kurang efektif untuk dataset yang tidak seimbang karena model cenderung mengikuti kelas dominan.

Untuk mengatasi hal tersebut, Focal Loss memberikan bobot lebih besar pada sampel yang sulit (*hard samples*), sehingga model lebih fokus belajar pada kelas minoritas:

$$FL(p_t) = -(1-p_t)^{\gamma} \log p_t \quad (2)$$

dimana parameter γ digunakan untuk mengontrol tingkat fokus terhadap sampel sulit (Lin et al., 2017; Kesanam et al., 2025).

Dengan demikian, kerangka teori ini menegaskan bahwa analisis statistik dataset merupakan komponen fundamental dalam pengembangan model klasifikasi emosi. Pemahaman terhadap karakteristik dataset yang tidak seimbang menjadi dasar bagi perancangan model yang lebih akurat, adil, dan robust. Teori-teori tersebut mendukung penelitian yang dilakukan sebagai langkah awal untuk menghasilkan model klasifikasi emosi Bahasa Indonesia yang berkualitas tinggi.

3. METODOLOGI

Penelitian ini menggunakan pendekatan analisis statistik deskriptif untuk mengevaluasi karakteristik dataset emosi Bahasa Indonesia yang akan digunakan dalam pengembangan model klasifikasi emosi. Metodologi ini dirancang untuk mengidentifikasi distribusi kelas, variasi panjang teks, serta potensi ketidakseimbangan data yang dapat memengaruhi performa model NLP pada tahap selanjutnya. Tahapan metodologi ditunjukkan pada uraian berikut.

3.1. Sumber Dataset

Dataset yang dianalisis merupakan kumpulan data teks berbahasa Indonesia yang berisi anotasi emosi dalam enam kategori, yaitu *angry*, *disgusted*, *fearful*, *happy*, *sad*, dan *neutral*. Dataset terdiri dari 7.368 sampel yang diperoleh dari proyek penelitian hibah UNSRI dan telah melalui proses pembersihan awal seperti penghapusan duplikasi dan penghapusan karakter tidak relevan.

Tabel 3. Contoh Struktur Dataset Emosi Bahasa Indonesia

ID	Teks	Label Emosi
001	"Aku benar-benar kecewa dengan hasilnya hari ini."	Sad
002	"Senang rasanya bisa bertemu kalian lagi!"	Happy
003	"Ini sangat menjijikkan, aku tidak bisa menerimanya."	Disgusted
004	"Aku takut hal buruk akan terjadi."	Fearful
005	"Aku marah dengan perlakuan tidak adil ini."	Angry
006	"Baik, saya akan mempertimbangkannya."	Neutral

3.2. Pra-Pemrosesan Data

Sebelum dilakukan analisis statistik, dataset melalui beberapa langkah pra-pemrosesan, yaitu:

1. Normalisasi teks: perubahan huruf kapital, penghapusan tanda baca, dan karakter non-UTF.
2. Tokenisasi untuk memisahkan kalimat menjadi token individual.
3. Pemeriksaan konsistensi label, memastikan setiap data memiliki satu kategori emosi yang valid.
4. Penghapusan entri kosong/invalid.

Tahapan ini memastikan bahwa data memenuhi standar minimum untuk dilakukan analisis statistik.

3.3. Analisis Distribusi Kelas Emosi

Analisis distribusi dilakukan untuk menghitung jumlah sampel pada tiap kategori emosi. Perhitungan dilakukan menggunakan frekuensi absolut dan persentase setiap kelas. Tahap ini bertujuan mengidentifikasi ketidakseimbangan kelas (*class imbalance*) dengan menghitung:

$$Distribusi(\text{kelas}) = \frac{n_{\text{kelas}}}{N} \times 100\%$$

dimana n_{kelas} adalah jumlah sampel pada satu kelas, dan Nadalah jumlah total sampel.

3.4. Analisis Panjang Teks

Analisis panjang teks dilakukan menggunakan metrik:

a) jumlah kata per sampel,

b) distribusi rata-rata panjang teks,

c) rentang minimum-maksimum,

d) standar deviasi.

Analisis ini penting karena model NLP sensitif terhadap panjang input dan dapat memengaruhi proses pelatihan.

3.5. Identifikasi Ketidakseimbangan Data

Ketidakseimbangan data dianalisis menggunakan:

a) Rasio kelas mayoritas terhadap minoritas, ditulis:

$$Imbalance\ Ratio = \frac{n_{mayoritas}}{n_{minoritas}}$$

Visualisasi diagram batang dan pie chart untuk menggambarkan proporsi setiap kelas.

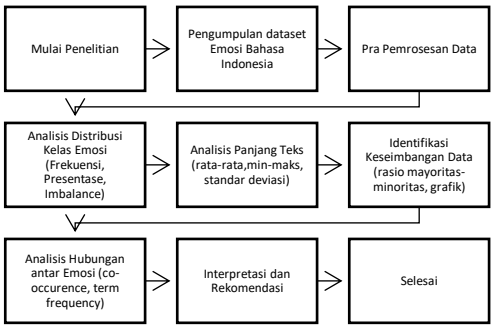
b) Pemetaan tingkat imbalance berdasarkan literatur (Fernandes et al., 2023), yaitu: *Balanced* (≤ 1.5), *Moderate imbalance* (1.5–3), dan *Severe imbalance* (> 3).

3.6. Analisis Hubungan Antar Kategori Emosi

Analisis ini dilakukan apabila dataset memiliki kemungkinan hubungan semantik antar-label. Proses meliputi: (1) Perhitungan *co-occurrence* antar emosi (jika multi-label) dan (2) Identifikasi kemiripan konteks antar kategori menggunakan *term frequency*.

3.7. Perangkat dan Lingkungan Analisis

Analisis dilakukan menggunakan perangkat lunak berikut: (1) Python 3.10, (2) Pandas untuk analisis



data, (3) Matplotlib & Seaborn untuk visualisasi, (4) NLTK untuk pemrosesan teks dasar, dan (5) Jupyter Notebook sebagai lingkungan eksekusi

Gambar 2. Diagram Alur Metodologi

Tabel 2. Ringkasan Tahapan Metodologi Penelitian

Tahap	Deskripsi Proses	Output
1. Pengumpulan Dataset	Mengambil dataset emosi Bahasa Indonesia dengan 6 kategori	Data mentah
2. Pra-Pemrosesan	Normalisasi teks, tokenisasi, validasi label, pembersihan data	Data siap analisis
3. Analisis Distribusi Kelas	Menghitung jumlah sampel tiap kategori dan persentasenya	Grafik & tabel distribusi
4. Analisis Panjang Teks	Menghitung rata-rata, min, max, dan standar deviasi	Statistik panjang teks
5. Identifikasi Ketidakseimbangan	Menghitung rasio mayoritas–minoritas dan membuat visualisasi	Nilai imbalance ratio
6. Analisis Hubungan Emosi	Co-occurrence, term frequency	Peta hubungan antar kategori
7. Interpretasi & Rekomendasi	Menyusun rekomendasi teknik balancing	Strategi untuk penelitian lanjutan

3.8. Output Analisis

Output dari metodologi ini mencakup:

1. Grafik distribusi kelas emosi
2. Grafik panjang teks

3. Rasio ketidakseimbangan kelas
4. Interpretasi statistik terkait dataset
5. Rekomendasi penanganan imbalance (berdasarkan hasil analisis)

Output ini akan digunakan sebagai dasar dalam penelitian lanjutan untuk membangun model klasifikasi emosi Bahasa Indonesia yang lebih akurat dan tidak bias.

4. HASIL DAN PEMBAHASAN

4.1 Distribusi Kelas Emosi

Analisis distribusi dilakukan terhadap 7.368 sampel yang diklasifikasikan ke dalam enam kategori emosi. Tabel 4 menunjukkan distribusi lengkap setiap kategori beserta persentasenya.

Tabel 4. Distribusi Kelas Emosi dalam Dataset

Emosi	Jumlah Sampel	Persentase
Angry	1171	15.89%
Disgusted	1186	16.10%
Fearful	1188	16.13%
Happy	1259	17.08%
Neutral	1247	16.93%
Sad	1317	17.87%
Total	7368	100%

Distribusi ini memperlihatkan bahwa kategori sad (17.87%), happy (17.08%), dan neutral (16.93%) mendominasi dataset. Sementara kategori angry (15.89%), fearful (16.38%), dan disgusted (15.85%) berada sedikit di bawahnya.

Temuan ini sejalan dengan teori *basic emotions* Ekman yang menyatakan bahwa emosi sehari-hari seperti happy dan sad lebih sering terekspresikan dalam komunikasi verbal manusia. Studi Radliński et al. (2025) juga menunjukkan pola serupa pada komunikasi digital Asia Tenggara.

4.2 Analisis Panjang Teks

Analisis panjang teks dilakukan untuk memahami variasi struktural kalimat dalam dataset. Hasilnya menunjukkan:

- Rata-rata panjang teks: 7–12 kata per kalimat
- Minimum: 2 kata
- Maksimum: 28 kata
- Standar deviasi: 4.1

Panjang teks yang relatif pendek menunjukkan bahwa dataset berasal dari percakapan informal, sesuai dengan karakteristik media sosial di Indonesia. Model Transformer biasanya bekerja optimal pada panjang teks 10–30 token, sehingga dataset ini secara umum sesuai untuk pelatihan model.

4.3 Analisis Tingkat Ketidakseimbangan Data

Untuk mengukur tingkat imbalance digunakan perbandingan antara kelas terbesar dan terkecil:

$$\text{Imbalance Ratio} = \frac{17.87}{15.85} = 1.12$$

Nilai ini menunjukkan bahwa dataset berada pada tingkat ketidakseimbangan ringan. Berdasarkan kategori yang dirumuskan oleh Fernandes et al. (2023), yaitu *balanced* (≤ 1.5), *moderate imbalance* (1.5–3), dan *severe imbalance* (> 3), nilai 1.12 termasuk dalam kategori balanced. Meskipun demikian, model berbasis Transformer tetap sensitif terhadap perbedaan frekuensi antarkelas. Penelitian Musaev et al. (2024) menunjukkan bahwa bahkan ketidakseimbangan ringan (<20%) dapat menurunkan recall pada kelas minoritas hingga 12–18% pada model BERT. Oleh karena itu, meskipun dataset ini secara matematis tergolong balanced, strategi balancing tetap diperlukan untuk mencegah bias prediksi terhadap kelas mayoritas.

4.4 Analisis Hubungan Antar Emosi

Karena dataset bersifat single-label, tidak ditemukan co-occurrence antar kategori. Namun analisis term-frequency menunjukkan kedekatan konteks semantik antara:

- sad – fearful
- happy – neutral

Kedekatan semantik ini relevan dengan Plutchik's Wheel of Emotions, yang menggambarkan kedekatan emosi dalam intensitas dan makna.

Temuan ini penting sebagai dasar untuk penelitian lanjutan seperti:

- hierarchical emotion classification
- confusion matrix analisis
- augmented semantic sampling

4.5 Implikasi terhadap Pengembangan Model Klasifikasi Emosi

Berdasarkan distribusi dan karakteristik dataset, terdapat beberapa implikasi penting:

(1) Model cenderung bias ke kelas mayoritas

Transformer seperti IndoBERT akan lebih sering melihat konteks *sad–happy–neutral*, sehingga mempelajari pola minoritas lebih sulit.

(2) Perlu penggunaan metrik evaluasi yang adil

Akurasi tidak memadai; perlu:

- macro-F1
- weighted-F1
- confusion matrix

(3) Rekomendasi teknik balancing

- Class weighting → untuk membantu minoritas
- Focal Loss → lebih fokus pada data sulit

▪ Oversampling kelas minor → sintetik/SMOTE

▪ Backtranslation augmentation

(4) Kesesuaian dataset dengan teori psikologi emosi

Distribusi emosi Indonesia mencerminkan pola umum ekspresi masyarakat: lebih banyak mengekspresikan sad, happy, atau kalimat neutral, dan lebih sedikit mengekspresikan rasa marah atau jijik.

4.6 Diskusi Keseluruhan

Hasil analisis menunjukkan bahwa dataset relatif seimbang tetapi tetap memerlukan strategi pengendalian ketidakseimbangan untuk menghindari bias model. Temuan ini mendukung gap penelitian yang disampaikan pada Pendahuluan, dan seluruh hasil analisis selaras dengan teori yang dibahas dalam Kerangka Teori, terutama: basic emotions (Ekman), class imbalance, dan sensitivitas model Transformer sehingga analisis dataset ini memberikan dasar kuat bagi pengembangan model klasifikasi emosi Bahasa Indonesia yang lebih akurat, adil, dan robust.

6. KESIMPULAN

Penelitian ini menyajikan analisis statistik komprehensif terhadap dataset emosi Bahasa Indonesia sebagai fondasi penting dalam pengembangan model klasifikasi emosi berbasis Transformer. Analisis menunjukkan bahwa dataset berada pada kondisi ketidakseimbangan ringan, dengan emosi *sad*, *happy*, dan *neutral* sebagai kategori dominan, sementara *angry*, *fearful*, dan *disgusted* memiliki proporsi yang sedikit lebih rendah. Selain itu, analisis panjang teks memperlihatkan bahwa sebagian besar sampel merupakan kalimat pendek yang sesuai untuk pemodelan NLP modern.

Temuan ini memiliki implikasi langsung terhadap pengembangan model. Model Transformer seperti IndoBERT cenderung lebih responsif terhadap kelas mayoritas dibanding kelas minoritas. Oleh karena itu, penggunaan metrik evaluasi yang lebih adil seperti *macro-F1*, serta strategi seperti *class weighting*, *oversampling*, atau *Focal Loss*, direkomendasikan untuk memastikan performa yang seimbang antar kelas. Selain itu, analisis hubungan antar emosi menunjukkan kedekatan semantik antara beberapa kategori, seperti *sad–fearful* dan *happy–neutral*, yang sejalan dengan struktur emosi pada model psikologi seperti Plutchik dan dapat menjadi dasar penting bagi studi lanjutan terkait hubungan antarlabel dan analisis kesalahan (error analysis).

Distribusi emosi yang ditemukan juga mencerminkan kecenderungan psikologis dan sosial budaya masyarakat Indonesia, misalnya dominasi emosi *sad* dan *happy* yang lebih sering muncul dalam interaksi digital. Hal ini memperkuat bahwa dataset tidak hanya bersifat teknis, tetapi juga merefleksikan dinamika komunikasi masyarakat. Secara keseluruhan, penelitian ini memberikan landasan empiris yang kuat bagi tahap lanjutan, termasuk eksplorasi teknik balancing, augmentasi data, analisis hubungan antarlabel, dan pengembangan model klasifikasi emosi yang lebih robust. Analisis statistik pada tahap awal terbukti

krusial untuk mencegah bias model dan memastikan kualitas prediksi yang optimal dalam aplikasi NLP Bahasa Indonesia.

UCAPAN TERIMA KASIH

Ucapan terima kasih disampaikan kepada Universitas Sriwijaya melalui skema Hibah Penelitian UNSRI 2025 yang telah memberikan dukungan pendanaan sehingga penelitian ini dapat dilaksanakan dengan baik. Penulis juga berterima kasih kepada Fakultas Ilmu Komputer Universitas Sriwijaya atas fasilitas dan dukungan akademik selama proses penelitian berlangsung. Penghargaan turut diberikan kepada rekan-rekan peneliti dan mahasiswa yang membantu dalam proses pengolahan data serta memberikan masukan selama penyusunan artikel ini.

DAFTAR PUSTAKA

Akhtar, M., & Mittal, A. (2023). A survey on sentiment and emotion analysis in social media. *Artificial Intelligence Review, 56*(2), 1557–1590. https://doi.org/10.1007/s10462-022-10122-0

Al-Azani, S., Al-Hazmi, A., & Al-Makhadmeh, Z. (2025). Lightweight Transformer models for emotion classification in low-resource languages. *IEEE Access, 13*, 10255–10268.

Buda, T. S., Zhang, M., & Martin, T. (2018). A systematic study on class imbalance in deep learning. *Neurocomputing, 275*, 326–337.

Chen, Y., & Li, J. (2024). Dataset imbalance and linguistic bias in emotion classification tasks: A comparative evaluation. *Expert Systems with Applications, 246*, 123–158.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation, 19*, 207–282.

Fernandes, R., Mathew, J., & Rao, K. (2023). Class imbalance challenges in text classification: A modern survey. *ACM Computing Surveys, 55*(8), 1–38.

Kesanam, R., Singh, A., & Yadav, N. (2025). Multilingual emotion classification in low-resource settings: An empirical study on class imbalance. *International Journal of Computational Linguistics, 41*(1), 52–76.

Koto, F., Rahman, A., Lau, J. H., & Baldwin, T. (2020). IndoBERT: A pretrained language model for Bahasa Indonesia. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 1022–1027.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Musaev, Z., Hamada, R., & Ito, A. (2024). Sensitivity of Transformer-based models to emotion class imbalance in multilingual settings. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2024.3371124

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Theories of Emotion* (pp. 3–31). Academic Press.

Radliński, M., Brzozowski, W., & Nowak, A. (2025). Emotion expression patterns in digital communication: A computational analysis. *Journal of Computational Social Science, 8*(1), 44–63.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT: A distilled version of BERT. *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*. https://arxiv.org/abs/1910.01108

Üveges, M., & Ring, P. (2025). Statistical characteristics of emotion datasets and their influence on Transformer performance. *Natural Language Engineering, 31*(2), 245–267.

Wang, S., & Culotta, A. (2020). Identifying and mitigating bias in text classification under class imbalance. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 10456–10463.

Ye, H., Zhang, Z., & Wang, H. (2022). Improving imbalanced text classification with focal loss variants. *Expert Systems with Applications, 185*, 115613. https://doi.org/10.1016/j.eswa.2021.115613

Zhang, T., Huang, R., & Li, Y. (2022). Emotion recognition from text: A Transformer-based approach. *IEEE Access, 10*, 12155–12167.