Jurnal Otomasi Kontrol dan Instrumentasi Vol 18 (1), 2026 ISSN: 2085-2517
https://doi.org/10.5614/joki.2026.18.1.1 E-ISSN: 2460-6340

Pemetaan Emosi Publik seputar Pengumuman UTBK 2025 di Indonesia: Pendekatan Multi-Label dengan Kalibrasi Bertarget

Mapping Public Emotions Regarding the 2025 UTBK Announcement in Indonesia: A Multi-Label Approach with Targeted Calibration

[1]Rizki Yustisia Sari*), [1]Sabar, [1]Joni, [1]Mardhiyatna, [1]Gelard Unthirta Pratama, [1]Ronal, [1]Fadli Hastito

[1]Rekayasa Instrumentasi dan Automasi, Institut Teknologi Sumatera, 35365, Lampung Selatan, Indonesia

*) *corresponding email: rizki.sari@ia.itera.ac.id*

Abstrak

Penelitian ini memetakan emosi publik pada cuitan berbahasa Indonesia terkait pengumuman UTBK 2025 dengan klasifikasi multi-label. Masalah utama dalam klasifikasi emosi multi-label pada media sosial adalah ketidakseimbangan label ekstrem, *distribution shift* antara data latih dan aplikasi, serta sinyal leksikal yang tipis untuk emosi tertentu. Tujuannya membangun pemodelan emosi yang reliabel pada korpus media sosial yang *long-tail* sekaligus mendemonstrasikan praktik kalibrasi pasca pelatihan yang dapat digeneralisasi. Kebaruan terletak pada integrasi empat komponen: (1) kalibrasi posterior per label dengan *Platt scaling*, (2) per-label *thresholding* bertarget presisi yang dibekukan dari *development set*, (3) *rate-targeting* berbasis kuantil skor untuk menyelaraskan prevalensi prediksi dengan *base rate* domain, dan (4) *lexicon-aware boost* terbatas konteks disertai final *clamp*. *Pipeline* ini dirancang *model-agnostic* dan ringan. Penelitian menggunakan pendekatan kuantitatif eksperimental dengan memvariasikan komponen kalibrasi pasca pelatihan untuk mengukur dampaknya terhadap performa klasifikasi. Model IndoBERTweet dilatih dengan *BCEWithLogitsLoss* pada data beranotasi *manual*, lalu dikalibrasi dan dievaluasi pada *dev/test*. Hasil menunjukkan keseimbangan kinerja mikro dan makro, peningkatan deteksi label minoritas, serta pemetaan pada 3500 cuitan dengan distribusi prevalensi yang konsisten dengan teori Plutchik.

Kata Kunci: multi-label, emosi, kalibrasi, IndoBERTweet, rate-targeting*,* lexicon boost*, Indonesia.*

Abstract

This study maps public emotions in Indonesian-language tweets related to the 2025 UTBK announcement using multi-label emotion classification. The main challenges in multi-label emotion classification on social media include extreme label imbalance, distribution shift between training and application data, and weak lexical signals for specific emotions. This study aims to build a reliable emotion modeling framework for long-tail social media corpora while demonstrating generalizable post-training calibration practices. The novelty lies in the integration of four components: (1) per-label posterior calibration using Platt scaling, (2) precision-targeted per-label thresholding frozen from the development set, (3) score-quantile–based rate targeting to align predicted prevalence with domain-based rates, and (4) context-limited lexicon-aware boosting with a final clamp. The proposed pipeline is lightweight and model-agnostic. This research adopts a quantitative experimental approach by varying post-training calibration components to measure their impact on classification performance. An IndoBERTweet model is trained using BCEWithLogitsLoss on manually annotated data, then calibrated and evaluated on development and test sets. The results demonstrate balanced micro- and macro-level performance, improved detection of minority labels, and emotion mapping over 3,500 tweets with prevalence distributions consistent with Plutchik's theory of emotions.

Keywords: multi-label, emotion, calibration, IndoBERTweet, rate-targeting, lexicon boost, Indonesia.

Makalah diterima 29 September 2025 – makalah direvisi 13 November 2025 – disetujui 19 Desember 2026

Karya ini adalah naskah akses terbuka dengan lisensi CC BY-SA.

1 Pendahuluan

Pengumuman Ujian Tulis Berbasis Komputer (UTBK) secara rutin memicu lonjakan percakapan dan emosi publik di media sosial Indonesia. Emosi-emosi ini membentuk persepsi kolektif terhadap keadilan, transparansi, dan kesiapan ekosistem pendidikan tinggi.

Dalam kerangka teori emosi, Plutchik mendefinisikan delapan emosi dasar (*joy, trust, fear, surprise, sadness, disgust, anger, anticipation*). Emosi-emosi ini dapat muncul bersama (*co-occurrence*) dan membentuk emosi turunan. Implikasinya, analisis harus mampu menangkap multi-emosi per unggahan, bukan sekadar polaritas.

Operasionalisasi modern atas roda emosi Plutchik, termasuk visualisasi dan pemetaan *co-occurrence*, memberi landasan konseptual untuk evaluasi hasil pada korpus dunia nyata [1].

Natural Language Processing (NLP) adalah cabang kecerdasan buatan yang memungkinkan komputer memahami dan memproses bahasa manusia. Dalam dekade terakhir, arsitektur Transformer telah merevolusi NLP melalui mekanisme *self-attention* yang mampu menangkap relasi kontekstual antar kata secara efisien. BERT (*Bidirectional Encoder Representations from Transformers*) merupakan model *Transformer* yang dipra-latih pada korpus besar untuk mempelajari representasi bahasa umum, kemudian dapat di-fine-tune pada tugas spesifik dengan dataset lebih kecil, proses yang disebut *transfer learning*. *Fine-tuning* BERT melibatkan penyesuaian bobot model pra-latih menggunakan data berlabel domain spesifik, sehingga model dapat beradaptasi pada karakteristik tugas target sambil mempertahankan pengetahuan linguistik yang telah dipelajari [2].

Kemajuan NLP Indonesia beberapa tahun terakhir diperkuat oleh tumbuhnya ekosistem riset dan sumber daya lokal seperti dataset, benchmark, dan komunitas yang mendorong kinerja model untuk teks media sosial Indonesia [2],[3]. Ketersediaan dataset emosi berbahasa Indonesia yang bertipe multi-label/multikelas semakin memperkuat urgensi studi emosi pada konteks lokal, bukan sekadar sentimen biner.

Namun, kesenjangan metodologis masih tampak pada tahap pasca proses ketika model dipindah-terapkan ke korpus baru. Hal ini terjadi di bawah berbagai bentuk *distribution shift*. Studi dan *benchmark* terkini menunjukkan performa model sering turun pada data "*in-the-wild*" dan membutuhkan strategi adaptasi yang hati-hati [4],[5].

Kesenjangan utama yang diidentifikasi adalah sebagai berikut. Pertama, banyak studi klasifikasi emosi multi-label belum melakukan penalaan ambang per-label (*threshold tuning*) yang optimal. Akibatnya, keseimbangan performa antar label melemah, terutama pada label minoritas. Penelitian mutakhir menegaskan bahwa optimisasi ambang berdampak besar pada F1 multi-label [6].

Kedua, fenomena *label shift* atau *prevalence shift* kerap menggeser distribusi emosi antara data latih dan korpus target. Namun, belum banyak studi yang menerapkan strategi kalibrasi prevalensi secara sistematis [7],[8].

Ketiga, penanganan label minoritas ekstrem (misalnya *disgust* pada korpus emosi) masih menjadi tantangan. Metrik konvensional seperti F1 kurang sensitif pada label yang sangat jarang [9],[10].

Keempat, untuk bahasa Indonesia, belum ada studi yang mengkombinasikan strategi kalibrasi posterior, *threshold tuning, rate-targeting*, dan *lexicon boost* dalam satu *pipeline* pasca pelatihan yang ringan dan model-*agnostic*.

Kelima, dari perspektif domain aplikasi, penelitian klasifikasi emosi berbahasa Indonesia umumnya berfokus pada topik umum (politik, produk komersial, pandemi) atau korpus *benchmark* generik [2],[11],[12],[13]. Belum ada studi yang secara khusus memetakan emosi publik pada konteks pendidikan tinggi, khususnya seputar pengumuman UTBK/SNBP momen kritis yang mempengaruhi jutaan siswa dan keluarga di Indonesia setiap tahunnya. Pemahaman pola emosional pada domain ini memiliki nilai praktis tinggi bagi pemangku kepentingan pendidikan untuk merancang komunikasi yang lebih empatik, mengidentifikasi kelompok rentan (misal: siswa dengan beban finansial), dan mengevaluasi persepsi publik terhadap sistem seleksi nasional.

Berdasarkan lima kesenjangan metodologis dan domain tersebut, penelitian ini merumuskan tiga masalah penelitian utama: Masalah Penelitian 1: Bagaimana menangani ketidakseimbangan label ekstrem pada data *long-tail* dalam klasifikasi emosi multi-label, khususnya untuk label minoritas seperti *disgust* yang prevalensinya <3% ?

Masalah Penelitian 2: Bagaimana mengatasi *distribution shift* (label *shift*/*prevalence shift*) antara korpus pelatihan dan aplikasi tanpa memerlukan label target atau pelatihan ulang model?

Masalah Penelitian 3: Bagaimana meningkatkan deteksi emosi dengan sinyal leksikal tipis (seperti *surprise* dan *anticipation*) yang tidak selalu memiliki penanda kata eksplisit?

Literatur mutakhir menyoroti pentingnya kalibrasi dan penanganan *shift* secara *post-hoc* yang ringan. Pendekatan ini tidak perlu mengubah arsitektur model melalui teknik seperti *temperature scaling* dan metode kalibrasi multi-kelas yang efisien [14],[15].

Menjawab ketiga masalah penelitian tersebut, penelitian ini mengusulkan skema pasca proses terkalibrasi yang ringan dan siap pakai untuk korpus Indonesia berbasis media sosial. Meskipun teknik kalibrasi pasca pelatihan dan *threshold tuning* telah digunakan secara terpisah dalam literatur [6],[9],[10], *novelty* penelitian ini terletak pada kombinasi sistematis empat strategi dalam satu *pipeline* terintegrasi: (1) kalibrasi *posterior*

per-label (*Platt/Isotonic*), (2) *threshold tuning* bertarget presisi, (3) *rate-targeting* berbasis kuantil untuk mengatasi *label shift*, dan (4) *lexicon-aware boost* terbatas konteks untuk label minoritas ekstrem. *Pipeline* ini bersifat *model-agnostic*, ringan secara komputasi (tidak memerlukan pelatihan ulang), dan dirancang khusus untuk bahasa Indonesia dengan karakteristik *long-tail*. Sejauh pengetahuan penulis, penelitian terdahulu yang mengintegrasikan keempat komponen ini untuk tugas klasifikasi emosi multi-label pada cuitan berbahasa Indonesia masih terbatas.

Secara spesifik, keempat komponen pipeline tersebut dirancang untuk menjawab ketiga masalah penelitian yang telah dirumuskan. Untuk itu, penelitian ini menetapkan tiga tujuan penelitian yang berkorespondensi:

Tujuan Penelitian 1: Mengembangkan strategi kalibrasi pasca pelatihan bertarget untuk menangani ketidakseimbangan label ekstrem, khususnya melalui kombinasi *precision-target*, *rate-target*, dan *lexicon-aware boost* untuk label minoritas.

Tujuan Penelitian 2: Mendemonstrasikan efektivitas *rate-targeting* berbasis kuantil skor dalam mengatasi *distribution shift* tanpa memerlukan label target atau pelatihan ulang model.

Tujuan Penelitian 3: Memvalidasi bahwa model berbasis IndoBERTweet dengan kalibrasi pasca pelatihan mampu menangkap pola emosi konsisten dengan kerangka teoretis Plutchik pada korpus domain UTBK.

Skema ini mencakup tiga komponen utama. Pertama, penyetelan ambang per-label berbasis *development set* untuk memaksimalkan F1. Kedua, kalibrasi prevalensi agar distribusi prediksi selaras dengan korpus target di bawah *label shift*. Ketiga, penguatan leksikon berbatas kata guna menjaga sensitivitas label minoritas (mis. *disgust*), disertai aturan *neutral-exclusive* dan penanganan baris kosong.

Penulis menerapkan skema ini pada cuitan berbahasa Indonesia terkait pengumuman UTBK 2025 menggunakan model IndoBERTweet yang dikalibrasi dengan korpus berlabel manual. Evaluasi dilakukan terhadap keseimbangan antar label dan konsistensi pola *co-occurrence* emosi.

Dengan demikian, kontribusi makalah ini adalah sebagai berikut. Pertama, *pipeline* multi-label yang akurat namun praktis untuk bahasa Indonesia. Kedua, bukti bahwa kalibrasi sederhana meningkatkan reliabilitas pada korpus dunia nyata di bawah *shift*. Ketiga, peta emosi publik UTBK yang sejalan dengan teori Plutchik sebagai validasi eksternal [2],[11].

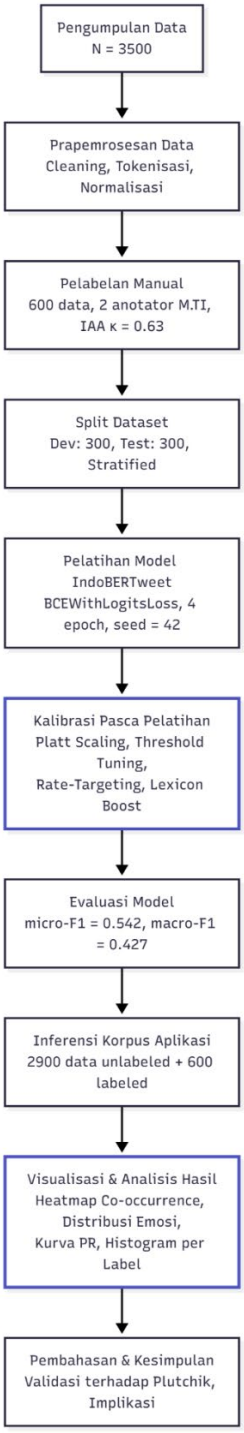
2 Metode

Penelitian ini menggunakan pendekatan kuantitatif eksperimental [16] berbasis teks untuk memetakan emosi publik dari cuitan berbahasa Indonesia pada platform X/Twitter. Desain eksperimental diterapkan dengan memanipulasi komponen kalibrasi pasca pelatihan (variabel independen: *Platt scaling*, *threshold tuning*, *rate-targeting*, *lexicon boost*) dan mengukur dampaknya terhadap performa klasifikasi multi-label (variabel dependen: *micro-F1*, *macro-F1*, *precision*, *recall*) dalam kondisi terkontrol dengan *seed* tetap dan pembagian *data stratified*.

Secara operasional, penulis melatih model *IndoBERTweet-base* pada korpus berlabel manual berisi 600 teks yang dianotasi oleh dua anotator independen. Korpus berlabel ini digunakan untuk kalibrasi dan penalaan ambang optimal per label. Model terkalibrasi kemudian diterapkan pada korpus aplikasi sebesar 3500 unggahan publik terkait UTBK 2025. Evaluasi berfokus pada keseimbangan performa antar label serta konsistensi pola *co-occurrence* emosi terhadap kerangka teori Plutchik.

Data diperoleh melalui proses pengambilan data menggunakan Twitter API, dengan bantuan Google Colab (Python). Pengambilan data dilakukan dengan menggunakan kata kunci yang relevan dengan UTBK/SNBP dalam rentang waktu sekitar pengumuman UTBK 2025 (D0), yaitu pada hari ke-8 setelah pengumuman (D0 + 8). Semua parameter pengambilan data tercatat dengan jelas dalam skrip yang digunakan, termasuk kata kunci yang digunakan untuk pencarian. Penanda waktu data dinormalisasi ke dalam Waktu Indonesia Barat (UTC+7) untuk menjaga konsistensi dan keseragaman waktu.

Tahap prapemrosesan mencakup serangkaian langkah, yaitu normalisasi teks, penghapusan URL, mention, dan retweet, deduplikasi dokumen, serta pemetaan emoji menjadi token khusus. Setelah seluruh tahapan prapemrosesan selesai, diperoleh sebanyak 3500 data yang telah dibersihkan dan siap untuk analisis lebih lanjut. Rangkaian tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Diagram alir tahapan penelitian

2.1 Pengumpulan Data

Pengumpulan data ulasan publik pada platform X/Twitter dilakukan menggunakan Twitter API v2 dengan kredensial resmi yang sah. Penulis menyusun kueri terstandar terkait pengumuman UTBK 2025, memfilter *tweet* berbahasa Indonesia, dan membatasi rentang waktu pengambilan pada jendela D0 hingga D0+8 hari yang berpusat pada tanggal resmi pengumuman (D0) untuk menangkap dinamika temporal di sekitar peristiwa tersebut [10],[11]. Daftar kata kunci yang digunakan meliputi "UTBK", "SNBP", "pengumuman UTBK", serta variasi ejaannya [14]. Desain jendela berbasis peristiwa dan kueri kata kunci ini mengikuti praktik sampling Twitter/X yang sah dan mempertimbangkan validitas analisis isi [15].

Proses pengambilan menghasilkan 3500 cuitan yang memenuhi kriteria kelayakan sebagai korpus penelitian. Setiap entri data memuat komponen utama berupa penanda waktu (*created_at*, *day_wib*, *dt_wib*), isi teks (*full_text*), bahasa (*lang*), identitas *tweet* (*id_str*, *conversation_id_str*), tautan (*tweet_url*), serta media (*image_url*, *jika ada*). Struktur percakapan terdokumentasi melalui *in_reply_to_screen_name* (jika merupakan balasan), sedangkan interaksi pengguna dicatat dalam bentuk metrik *favorite_count*, *retweet_count*, *reply_count*, dan *quote_count*.

Variabel identitas pengguna (seperti *user_id_str* dan *username*) disimpan untuk keperluan audit dan replikasi, namun tidak disertakan dalam publikasi naskah demi menjaga kerahasiaan dan untuk mematuhi kebijakan privasi yang berlaku. Seluruh cap waktu dinormalisasi ke zona UTC+7 (WIB) untuk konsistensi analisis. Dataset kemudian diarsipkan dalam format CSV, sehingga dapat direplikasi oleh peneliti lain. Cuplikan data pada Tabel 1 memberikan ilustrasi mengenai ragam komentar yang menjadi bahan analisis dalam penelitian ini.

Tabel 1. Contoh komentar yang dikumpulkan

No	Komentar	Tanggal
1	skrg aku prcaya kalo skor utbk>skor to ril adanya	28 Mei 2025
2	gua ciut bgt liat twit orang orang lulus utbk tapi kagak mampu bayar.. nasib gua gimana ya:(? gua memang masih 2 tahun lagi sih untuk kuliah tapi gua dari sekarang udah kepikiran huhuu	28 Mei 2025
3	eh sumpah ada ga si info loker yg dibayar harian daera bekasi-jakarta? gua butuh Ya Allah buat bayar mandiri capek bgt udah effort bljr buat utbk keterima malah ditolak mentah sm keluarga maunya yg deket aja kuliahnya. skrg juga ga mau biayain ARGHHH	28 Mei 2025
4	Life After ketolak utbk mikir gini. Mending yang penting kuliah that mean pokoknya lulus suka gak suka urusan belakangan kalau rejekinya disitu yaudah. Atau kuliah yang penting? Yang dimana lo kudu usahain bener bener jurusan apa yang lo mau bahkan kalau lo harus semi/gap year?	29 Mei 2025
5	Skor T0 tidak pernah mencapai 650. Skor SNBT 660. Tuhan Yesus baik! https://t.co/CQArbG8I65	30 Mei 2025
6	NAPA GUE JDI BIJAK. HIKMAH UTBK KE 20. ANJAY	1 Juni 2025
7	@crocodilaa47 hii gapapa kokk pasti ada jalan yg lebihh indaaah disanaa trust in God sayangg. kamu udah ngerasain utbk tahun ini tahun depan pasti bakalan lebih gacorrrr lagii semangat orang hebatttt!!	3 Juni 2025
8	tidur dengan nyenyak karna lolos snbt	29 Mei 2025
9	GAK NYANGKA SAMPE SEKARANG AKUTUH LOLOS UTBK UHUYYY BISA KETERIMA N KULIAH DI BALI UDAH ANOTHER LEVEL OF HAPPINESS Btw yang tau info kost di bali https://t.co/oUWNhuzUJa	6 Juni 2025
10	Gue gak lolos utbk anjingggg	28 Mei 2025

2.2 Prapemrosesan Data

Prapemrosesan teks dilakukan dengan Python menggunakan pustaka *pandas*, *re*, dan *nltk* untuk menyiapkan korpus pemetaan emosi. Prosedur ini menyingkirkan elemen non-leksikal, menata struktur teks, dan meningkatkan akurasi algoritma klasifikasi emosi. Langkah-langkah Prapemrosesan yang diterapkan meliputi:

1. Normalisasi Huruf
   Seluruh teks diubah menjadi huruf kecil untuk menyamakan representasi kata dan menghindari duplikasi analisis akibat variasi kapitalisasi (misalnya “UTBK” dan “utbk”).
2. Pembersihan URL, *mention*, dan *tag* “RT”
   Elemen non-leksikal seperti tautan, akun pengguna, serta tag *retweet* dihapus guna menjaga fokus analisis pada isi teks yang bermakna.

3. Penghapusan tanda baca, angka, dan simbol
   Karakter non-alfabet dibuang agar mengurangi *noise* dan menekankan kata-kata utama yang relevan dengan emosi pengguna.
4. Pemetaan emoji ke token emosi
   Emoji ditransformasikan menjadi token khusus yang merepresentasikan efek tertentu, sehingga ekspresi emosional tetap terjaga dalam korpus.
5. Tokenisasi
   Teks diuraikan menjadi satuan token (umumnya kata/tanda baca) agar tiap unit dapat diproses secara mandiri oleh algoritma NLP, langkah ini membantu pemetaan emosi pada tingkat kata.
6. S*topword removal*
   Kata-kata fungsional berdaya informasi rendah (mis. “yang”, “dan”, “dari”) dihapus menggunakan daftar *stopword* dari pustaka NLTK guna mengurangi *noise* dan menonjolkan fitur yang relevan secara semantik.
7. *Lemmatization / Stemming*
   Kata dikembalikan ke bentuk dasar atau lema agar variasi morfologis tidak menimbulkan redundansi fitur dan model lebih efisien dalam mengenali pola emosi.
8. Deduplikasi
   Cuitan yang identik atau sangat mirip (berdasarkan ID dan *hashtag*) dihapus untuk memastikan tidak ada bias akibat pengulangan data.

Tabel 2. Contoh komentar hasil prapemrosesan

No	Komentar	Tanggal
1	sekarang percaya skor utbk skor to real ada	28 Mei 2025
2	gua ciut lihat orang lulus utbk tidak mampu bayar nasib gua bagaimana gua masih dua tahun lagi kuliah gua sekarang sudah pikir	28 Mei 2025
3	sumpah ada tidak info loker bayar harian daerah bekasi jakarta gua butuh allah bayar mandiri capek sudah belajar utbk diterima malah tolak keluarga maunya dekat kuliah sekarang juga tidak mau biayai	28 Mei 2025
4	ketolak utbk pikir mending penting kuliah pokok lulus suka tidak suka urusan belakangan kalau rejeki disitu ya sudah atau kuliah penting dimana lo harus usah benar jurusan mau bahkan kalau harus gap year	29 Mei 2025
5	skor to tidak pernah capai skor snbt tuhan yesus baik	30 Mei 2025
6	kenapa gue jadi bijak hikmah utbk anjay	1 Juni 2025
7	gapapa pasti ada jalan lebih indah disana percaya tuhan sayang kamu sudah rasa utbk tahun ini tahun depan pasti bakal lebih semangat orang hebat	3 Juni 2025
8	tidur nyenyak karena lolos snbt	29 Mei 2025
9	tidak sangka sampai sekarang aku lolos utbk bisa terima kuliah bali sudah level bahagia yang tahu info kost bali	6 Juni 2025
10	gue tidak lolos utbk anjing	28 Mei 2025

2.3 Pembagian Set dan Komposisi Label

Dari total 3500 data hasil prapemrosesan, 600 data dipilih secara acak untuk dianotasi manual oleh dua anotator independen berkualifikasi Magister Teknologi Informasi dari Fasilkom Universitas Indonesia guna membentuk korpus berlabel *manual* (*manually annotated corpus*). Korpus berlabel ini kemudian dibagi menjadi *development set* (*dev*, n=300) dan *test set* (*test*, n=300) untuk keperluan kalibrasi dan evaluasi akhir. Sisa 2900 data tidak berlabel digunakan sebagai korpus aplikasi untuk pemetaan emosi.

Ukuran korpus berlabel (600 sampel) memang berada di bawah standar ideal 1000+ sampel untuk *fine-tuning* BERT pada tugas klasifikasi teks. Namun, beberapa pertimbangan membenarkan konfigurasi ini: (1)

IndoBERTweet sudah dipra latih pada korpus Twitter Indonesia berukuran besar (>200 juta *tweet*), sehingga telah mempelajari representasi domain-spesifik yang kuat; (2) Fokus penelitian adalah kalibrasi pasca pelatihan, bukan optimasi *fine-tuning*, 600 sampel cukup untuk mendemonstrasikan efektivitas strategi kalibrasi pada korpus terbatas; (3) Pembagian 50-50 (*dev-test*) dipilih untuk memaksimalkan ukuran *test* set guna evaluasi *robust*, mengikuti praktik pada dataset emosi berukuran kecil-menengah; (4) Risiko *overfitting* dimitigasi melalui *early stopping*, *dropout*, dan validasi pada *dev set* terpisah. Meski demikian, kami mengakui keterbatasan ini dan merekomendasikan perluasan korpus berlabel pada studi lanjutan. Kedua anotator merupakan penutur asli bahasa Indonesia dengan pengalaman dalam NLP dan dipilih berdasarkan pemahaman mendalam terhadap konsep emosi serta pengalaman praktis dalam anotasi teks, untuk memastikan kualitas korpus yang lebih tinggi dibanding anotasi oleh anotator *non-expert*.

Pemisahan dilakukan secara acak dengan *seed* tetap (*seed* = 42) guna menjamin keterulangan hasil (*reproducibility*). *Split* dilakukan dengan *stratified sampling* untuk mempertahankan distribusi proporsional setiap label emosi di *dev* dan *test*, menggunakan fungsi *train_test_split* dari *scikit-learn* dengan parameter *stratify* berdasarkan label mayoritas per sampel. Strategi ini penting untuk mencegah bias evaluasi akibat ketidakseimbangan label ekstrem.

Tabel 3. Statistik deskriptif korpus berlabel manual (600 data)

Label	Jumlah sampel positif	Persentase	*Support dev* (n=300)	*Support test* (n=300)
Neutral	376	62.7%	182	194
Surprise	71	11.8%	39	32
Joy	65	10.8%	31	34
Anticipation	32	5.3%	16	16
Trust	30	5.0%	14	16
Sadness	29	4.8%	10	19
Anger	26	4.3%	14	12
Fear	21	3.5%	11	10
Disgust	13	2.2%	8	5
Total	663	110.5%	325	338

Catatan: Total persentase >100% karena skema multi-label (satu sampel bisa memiliki beberapa label).

Berdasarkan Tabel 3, total penetapan label berjumlah 663 pada 600 sampel berlabel, sehingga label *cardinality* sebesar 663/600 = 1.11. Dengan total kemungkinan 9 label per sampel, label *density* diperoleh sebagai 663/(600×9) = 0.123. Nilai ini menunjukkan bahwa, rata-rata, setiap sampel memuat sekitar 1 label dan hanya sekitar 12.3% dari ruang label yang terisi, merefleksikan karakteristik *long-tail* pada korpus. Perlu dicatat bahwa persentase per label dapat berjumlah lebih dari 100% karena sifat tugas multi-label yang memungkinkan banyak label benar per sampel. Selain itu, *disgust* muncul hanya pada 2.2% sampel (13/600), menegaskan ketidakseimbangan yang tinggi pada label minoritas.

2.4 Pelabelan Manual dan Kualitas Korpus

Panduan anotasi mendeskripsikan sembilan label emosi (*anger, anticipation, disgust, fear, joy, sadness, surprise, trust,* dan *neutral*) berdasarkan kerangka Plutchik, disertai contoh positif dan negatif, aturan *neutral-exclusive* untuk kasus tanpa indikasi emosi, serta penanganan ambiguitas umum pada bahasa Indonesia di media sosial. Pedoman memuat kriteria inklusi dan eksklusi per label, prioritas penentuan label ketika beberapa emosi muncul bersamaan, dan daftar *edge cases* yang sering menimbulkan tafsir ganda.

Kedua anotator bekerja secara independen tanpa komunikasi selama fase pelabelan. Sebelum anotasi utama, keduanya mengikuti pelatihan terstruktur selama empat jam yang mencakup pengenalan delapan emosi dasar Plutchik dan karakteristiknya, latihan identifikasi emosi pada contoh cuitan, penanganan kasus multi-label dan ambiguitas, serta *pilot annotation* pada 50 sampel untuk kalibrasi pemahaman. Setiap teks dapat menerima 0 hingga 9 label emosi sesuai skema multi-label. Ketidaksepakatan diselesaikan melalui adjudication oleh peneliti senior dengan mekanisme diskusi konsensus. Untuk kasus ambiguitas tinggi, digunakan voting mayoritas setelah meninjau pedoman dan contoh pembanding.

Tingkat kesepakatan antar-anotator diukur menggunakan Cohen's κ per label pada 600 contoh beranotasi ganda sebelum adjudication. Hasilnya menunjukkan rata-rata κ = 0.63 dengan rentang 0.46-0.80 (*neutral*

0.80; *sadness* 0.73; *joy* 0.70; *anger* 0.66; *trust* 0.62; *fear* 0.59; *anticipation* 0.56; *surprise* 0.53; *disgust* 0,46). Secara keseluruhan, nilai ini berada pada kategori substantial menurut skala Landis & Koch. Label dengan penanda leksikal lebih eksplisit (mis. *neutral, sadness, joy*) cenderung memiliki κ lebih tinggi, sedangkan label yang lebih kontekstual atau jarang (mis. *surprise, anticipation, disgust*) berada pada kisaran moderate. Walaupun para anotator bukan spesialis psikologi, pelatihan berbasis kerangka teoretis dan capaian IAA yang acceptable menunjukkan kualitas anotasi yang memadai untuk tugas klasifikasi emosi di media sosial, selaras dengan praktik komunitas NLP bahwa anotator non-expert dengan pelatihan terstruktur dapat menghasilkan label yang *reliable* [16],[17].

2.5 Uji Reliabilitas Anotasi (Uji IAA: *Cohen's Kappa* per-label)

Reliabilitas anotasi dinilai dengan *inter-annotator agreement* (IAA) menggunakan Cohen's κ per-label dan F1 antar-anotator pada 600 sampel sebelum *adjudication*. Metrik ini menunjukkan konsistensi penilaian antara kedua anotator independen.

Tabel 4 menunjukkan hasil perhitungan *Cohen's Kappa* untuk setiap label emosi sebelum proses *adjudication*.

Tabel 4 *Inter-Annotator Agreement* per Label (N=600, sebelum *adjudication*)

Label	*Cohen's Kappa*	Interpretasi	*Agreement* (%)	F1 antar-anotator	*Disagreement Cases*
Neutral	0.80	*Substantial*	91.7%	0.94	50
Sadness	0.73	*Substantial*	98.0%	0.74	12
Joy	0.70	*Substantial*	95.3%	0.73	28
Anger	0.66	*Substantial*	97.8%	0.67	13
Trust	0.62	*Substantial*	97.3%	0.64	16
Fear	0.59	*Moderate*	98.0%	0.60	12
Anticipation	0.56	*Moderate*	96.8%	0.58	19
Surprise	0.53	*Moderate*	92.8%	0.57	43
Disgust	0.46	*Moderate*	98.5%	0.47	9
Rata-rata	0.63	*Substantial*	96.3%	0.66	22

Interpretasi mengikuti skala Landis & Koch (1977): <0.00 *poor,* 0.00-0.20 *slight,* 0.21-0.40 *fair,* 0.41-0.60 *moderate,* 0.61-0.80 *substantial,* 0.81-1.00 *almost perfect.*

Rata-rata *Cohen's* κ = 0.63 menunjukkan tingkat kesepakatan substantial secara keseluruhan. Label dengan penanda leksikal yang lebih eksplisit (*neutral* κ=0.80, *sadness* κ=0.73, dan *joy* κ=0.70) mencapai κ ≥ 0.70, sedangkan label dengan ambiguitas semantik lebih tinggi seperti *disgust* (κ=0.46), *surprise* (κ=0.53), dan *anticipation* (κ=0.56) berada di rentang *moderate*. *Fear* (κ=0.59) juga mendekati batas *moderate*. Pola ini konsisten dengan literatur bahwa emosi dasar yang diekspresikan secara jelas lebih mudah dianotasi secara konsisten dibanding emosi yang bergantung konteks [16],[17].

Rata-rata terdapat 22 kasus ketidaksepakatan per label yang kemudian diselesaikan melalui *adjudication* oleh *resolver* ketiga lewat diskusi konsensus. Ketidaksepakatan paling sering muncul pada: (1) ambiguitas *neutral* vs. *anticipation*, ketika tuturan informatif tentang rencana sulit dibedakan; (2) *overlap sadness* vs. *fear*, misalnya ekspresi kecemasan finansial atau akademik bisa diinterpretasikan sebagai kesedihan atau ketakutan; dan (3) multi-label *edge cases*, yaitu perbedaan penilaian apakah tuturan memuat dua emosi sekaligus atau hanya satu emosi berintensitas tinggi.

2.6 Metrik Evaluasi

Evaluasi performa model menggunakan metrik standar untuk klasifikasi multi-label:

1. *Micro-F1*: Menghitung F1 secara agregat dengan menjumlahkan *true positive, false positive,* dan *false negative* dari semua label. Lebih sensitif terhadap performa pada label mayoritas.
2. *Macro-F1*: Menghitung F1 per-label kemudian dirata-ratakan. Memberikan bobot yang sama pada setiap label, sehingga lebih sensitif terhadap performa pada label minoritas [18].
3. *Precision* dan *Recall* (*Micro* & *Macro*): Mengukur ketepatan dan kelengkapan prediksi.
4. *Hamming Loss*: Proporsi label yang salah diprediksi per sampel (semakin kecil semakin baik).
5. ROC-AUC per label: Mengukur kemampuan diskriminatif model untuk setiap label.
6. *Cohen's Kappa* per-label: Mengukur kesepakatan antar-anotator.

2.7 Pelatihan Model (Pelatihan IndoBERTweet)

Pelatihan menggunakan IndoBERTweet (*indolem/indobertweet-base-uncased*) sebagai model Transformer untuk klasifikasi multi-label sembilan emosi (*anger, anticipation, disgust, fear, joy, sadness, surprise, trust, neutral*).

Penulis menggunakan IndoBERTweet-base sebagai *Transformer backbone* (12 *layer*, ukuran *hidden* 768), diikuti sebuah *linear classification head* (*fully connected*) dengan 9 unit keluaran, masing-masing mewakili satu label emosi. Setiap unit keluaran dipasangi sigmoid untuk menghasilkan probabilitas independen per label sesuai skema klasifikasi multi-label. Tidak ada komponen arsitektural tambahan (misal: *task-specific attention* atau *label dependency modeling*) karena fokus penelitian berada pada kalibrasi pasca pelatihan, bukan modifikasi arsitektur.

Tokenisasi mengikuti tokenizer IndoBERTweet dengan panjang maksimum 160 token, menerapkan *padding* dan *truncation* pada *max_length* serta *lowercasing* bawaan. Pelatihan dilakukan dengan PyTorch/Transformers memakai AdamW (*learning rate* 2×10^{-5}, *weight decay* 0.01) dan penjadwalan *linear warm-up* 10% dari total langkah. *Loss function* yang digunakan adalah BCEWithLogitsLoss, *batch size* 16, 4 *epoch*, dan *seed* = 42 ditetapkan konsisten pada seluruh komponen relevan untuk memastikan keterulangan.

Evaluasi internal dijalankan berkala pada himpunan *dev. checkpoint* terbaik dipilih berdasarkan *micro-F1* (dengan *macro-F1* sebagai *tie breaker*). Konfigurasi terbaik yang diperoleh adalah *checkpoint*-616, yang kemudian digunakan untuk seluruh evaluasi akhir dan tahap aplikasi.

2.7.1 Kalibrasi pada *Multi-label Classification*

Berbeda dari *multi-class classification* yang memakai *softmax* dan memaksa tepat satu kelas aktif per sampel, multi-label *classification* memperlakukan setiap label secara independen dengan sigmoid pada tiap *output*. Konsekuensinya, kalibrasi juga dilakukan per label, bukan pada satu vektor probabilitas global.

Pada penelitian ini, penulis menerapkan *post-hoc calibration* per label di korpus berlabel (600 sampel) menggunakan *Platt scaling* dan *isotonic regression*.

1. *Platt scaling* memodelkan transformasi logistik pada logit/skor mentah s:

$$\hat{p} = \sigma(A\,s + B) = \frac{1}{1 + \exp\,(A\,s + B)} \tag{1}$$

2. *Isotonic regression* adalah alternatif non-parametrik yang memetakan skor ke probabilitas terkalibrasi melalui fungsi monoton naik 10.

Kedua metode lazim untuk kalibrasi multi-label karena ringan secara komputasi dan tidak memerlukan *retraining* model [6],[10].

Setelah kalibrasi, penulis menentukan *threshold* per label pada *development set* dengan target presisi tinggi (untuk menekan *false positive* pada label minoritas), lalu membekukan *threshold* tersebut sebelum evaluasi *ditest set*.

2.8 Evaluasi di *Dev*/Test

Tahap *dev* berfungsi ganda sebagai seleksi model dan penentuan ambang awal. Setelah setiap *epoch*, model dievaluasi pada *dev* (n = 300). *Micro-F1* digunakan sebagai kriteria utama pemilihan, dengan *macro-F1* sebagai *tie-breaker* untuk menjaga keseimbangan lintas label. Pada tahap ini, dilakukan penyapuan ambang per label di rentang 0.05–0.95 untuk memperoleh titik keputusan awal terbaik.

Evaluasi final dijalankan sekali pada test (n = 300) menggunakan *checkpoint*-616 dan set ambang hasil *dev* (tanpa penyesuaian ulang di *test*). Metrik yang dilaporkan mencakup *micro/macro-F1*, *precision/recall* (*micro* & *macro*), ROC-AUC per label, serta *Hamming loss*.

2.9 Kalibrasi Pasca Pelatihan

Model terbaik (*checkpoint*-616) dipilih melalui evaluasi berkala pada *development set* (*dev*) dengan kombinasi metrik *micro-F1* dan *macro-F1*. Praktik ini lazim pada *imbalanced dataset*, karena *micro-F1* lebih peka terhadap label mayoritas (berfrekuensi tinggi), sedangkan *macro-F1* menyeimbangkan kontribusi label minoritas (berfrekuensi rendah) [18]. Model IndoBERTweet, yang dipra-latih pada korpus cuitan berbahasa

Indonesia, memberikan representasi yang lebih selaras dengan ragam bahasa dan *slang* lokal dibandingkan model umum [2].

Evaluasi awal menunjukkan satu label emosi sama sekali tidak terdeteksi (F1 = 0). Untuk mengatasinya, penulis menerapkan *post-hoc calibration* yang tidak mengubah bobot model dan hanya menyetel parameter keputusan.

Komponen Kalibrasi (*post-hoc).*

1. Kalibrasi *posterior* menggunakan *Platt scaling* atau *isotonic regression* pada korpus berlabel guna memetakan skor mentah menjadi probabilitas terkalibrasi.
2. Kalibrasi ambang (*threshold calibration*) berbasis *precision-target* dan *rate-target* untuk menyesuaikan titik keputusan per label.
3. Penetapan ambang per label (termasuk ambang khusus untuk *disgust*), mengikuti rekomendasi kalibrasi multi-label [6],[7],[20].
4. Penguatan leksikon *(lexicon-aware boost*) terbatas konteks, disusul *final clamp* agar sensitivitas label minoritas meningkat tanpa menimbulkan prediksi berlebih.

Sebagai tambahan, penulis menerapkan aturan *neutral-exclusive* dan anti-kosong untuk mencegah keluaran kosong serta menjaga konsistensi prediksi.

Pendekatan *post-hoc* ini *model-agnostic* dan ringan karena tidak memerlukan pelatihan ulang [14],[15]. Dengan demikian, *pipeline* mudah diadaptasi pada korpus baru yang mengalami *distribution shift* sambil mempertahankan stabilitas dan interpretabilitas hasil.

2.10 Inferensi pada Korpus Gabungan 3500 data

Inferensi dijalankan pada 3500 teks domain target menggunakan *checkpoint*-616. Proses mengikuti setelan tokenisasi saat pelatihan (panjang maksimum 160 token dengan *padding* dan *truncation* pada *max_length*), serta *batch size* yang diatur agar efisien di CPU sekaligus menjaga determinisme. Skor per label kemudian dikonversi menjadi label biner menggunakan ambang per label yang dibekukan dari *development set*: umum 0.25, khusus *surprise* 0.23, dan ambang akhir *disgust* = 0.122 (ditetapkan melalui kombinasi *precision-target*, *rate-target*, *lexicon-aware boost* berbatas kata, serta *final clamp*).

Aturan *neutral-exclusive* memastikan *neutral* aktif hanya bila tidak ada emosi lain yang melewati ambang, sedangkan mekanisme anti-kosong mengaktifkan *neutral* ketika seluruh skor berada di bawah ambang sehingga tidak ada keluaran kosong. *Pipeline* dapat direplikasi penuh karena ambang per label, pola leksikon, *seed*, dan versi pustaka disertakan bersama model.

2.11 Pemetaan dan Analisis Emosi

Analisis *co-occurrence* antar emosi disajikan sebagai tabel berisi jumlah dan persentase pasangan emosi yang muncul bersamaan, dilengkapi visualisasi diagram batang beresolusi tinggi yang merangkum frekuensi ko-okurensi utama. Kinerja model pada berbagai ambang keputusan dilaporkan melalui kurva *Precision–Recall* rata-mikro beserta ringkasannya (misalnya *area* di bawah kurva dan titik operasi terpilih). Untuk memastikan keberterapan dan replikasi, naskah ini mengekspisitkan keluaran pasca proses berupa daftar ambang per label hasil kalibrasi, termasuk ambang akhir khusus *disgust*, serta aturan yang digunakan, mencakup *neutral-exclusive*, mekanisme anti-kosong, pola leksikon untuk *lexicon-aware boost*, dan parameter *rate-target*. Selain itu, disampaikan catatan rilis yang memuat hiperparameter pelatihan, setelan tokenisasi, *seed*, dan versi pustaka yang dipakai, sehingga inferensi pada korpus baru dapat direplikasi dengan ambang dan aturan yang sama.

3 Hasil dan Diskusi

3.1 Ringkasan Hasil dan Tata Cara Evaluasi

Secara agregat, performa terbaik tercatat pada *micro-F1* = 0.542 dan *macro-F1* = 0.427, *dengan precision-micro* = 0.510, *recall-micro* = 0.579, serta *Hamming loss* = 0.109. Selisih *micro-F1* yang lebih tinggi daripada *macro-F1* mengindikasikan model lebih efektif pada label mayoritas, sementara *macro-F1* menegaskan masih adanya disparitas pada label minoritas [18]. Praktik pelaporan PR/F1 multi-label mengikuti konvensi *dataset* emosi berbutir halus seperti GoEmotions, yang menekankan pelaporan per label untuk memantau stabilitas pada kategori bernuansa [20].

Kualitas label diverifikasi pada N = 600 melalui *inter-annotator agreement*. Rata-rata Cohen's κ = 0.63 (*Substantial*) dengan *agreement* = 96.3% dan F1 antar-anotator = 0.66. Label dengan penanda leksikal eksplisit, *neutral* (κ = 0.80), *sadness* (0.73), dan *joy* (0.70), menunjukkan *substantial*, sedangkan *surprise* (0.53), *anticipation* (0.56), *fear* (0.59), dan *disgust* (0.46) berada pada kisaran Moderate. Rata-rata terdapat 22 kasus ketidaksepakatan per label yang diselesaikan melalui *adjudication* konsensus. Profil ini konsisten dengan literatur bahwa emosi eksplisit cenderung lebih stabil dibanding kategori yang kontekstual/ambigus, dan menjelaskan variasi kinerja per label pada hasil model.

Seluruh evaluasi menggunakan aturan *neutral-exclusive*: *neutral* aktif hanya jika tidak ada emosi lain yang melewati ambang pada suatu sampel. Aturan ini mencegah inflasi *neutral*, memudahkan interpretasi multi-label, dan menjaga konsistensi *operating point* saat kalibrasi.

Matriks keseluruhan ditampilkan pada Tabel 5.

Tabel 5. Matriks hasil keseluruhan

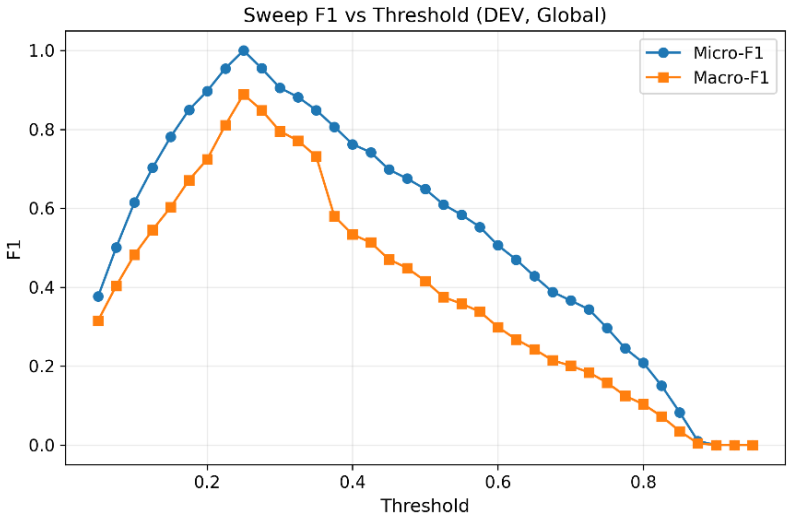
Threshold	*Precision (Micro)*	*Recall (Micro)*	*F1 (Micro)*	*Precision (Macro)*	*Recall (Macro)*	*F1 (Macro)*	*Hamming Loss*
0.25	0.5101	0.5790	0.5424	0.4266	0.4673	0.4272	0.1086

3.2 Sensitivitas Ambang dan Implikasinya

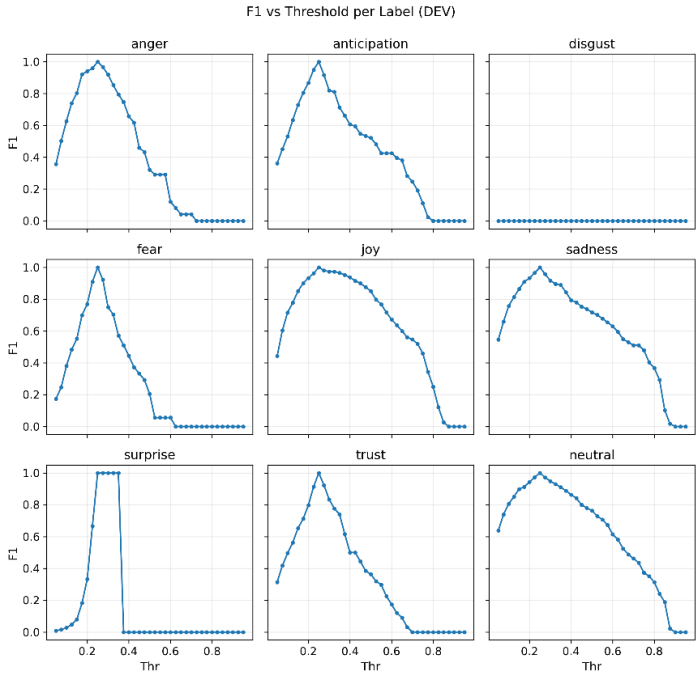
Kurva F1 terhadap *threshold* pada data *dev* menunjukkan rentang optimum sekitar 0.20–0.30, dengan puncak stabil pada angka 0.25 untuk *micro-F1*. Pada area *threshold* yang terlalu rendah, recall meningkat tetapi precision turun tajam; sebaliknya, pada *threshold* yang terlalu tinggi, precision naik sementara recall tereduksi sehingga banyak label positif tidak terdeteksi. Pola ini konsisten dengan karakter kalibrasi pasca pelatihan pada model modern serta perilaku umum pada *imbalanced dataset* [6], [7].

Berdasarkan temuan tersebut, penulis memilih ambang global 0.25 sebagai nilai dasar, kemudian menyetel ambang per label setelah *post-hoc calibration* (*Platt/isotonic*) di *dev* untuk mengimbangi perbedaan *base rate* dan kejelasan leksikal tiap emosi. Secara khusus, *surprise* = 0.23 dan *disgust* = 0.122, yang diturunkan melalui kombinasi *precision-target*, rate-target, *lexicon-aware boost* berbatas kata, serta *final clamp* [6], [7], [20]. Seluruh *threshold* dibekukan sebelum evaluasi pada *test* guna mencegah *over-tuning*.

Sebagai kontrol konsistensi, penulis menerapkan *neutral-exclusive* (*neutral* aktif hanya ketika tidak ada emosi lain melewati ambang) dan anti-kosong (mengaktifkan *neutral* ketika seluruh skor berada di bawah ambang). Kombinasi aturan ini efektif menekan keluaran kosong sekaligus mempertahankan interpretabilitas hasil, sejalan dengan praktik pelaporan PR/F1 per label pada studi emosi berbutir halus [20].



Gambar 2. Kurva *micro-F1* dan *macro-F1* terhadap *threshold* (*dev*, ambang global)



Gambar 3. Kurva F1 terhadap ambang probabilitas per label (*dev*)

Secara per-label, beberapa emosi tampil menonjol: *neutral* memperoleh F1 0.642 (*precision* 0.599; *recall* 0.693; AUC 0.834), *sadness* 0.622 (0.531; 0.750; AUC 0.906), *joy* 0.615 (0.543; 0.710; AUC 0.911), dan *trust* 0.563 (0.493; 0.655; AUC 0.905). Keempatnya umumnya hadir dengan penanda kata yang eksplisit atau pola pragmatik yang relatif konsisten, sehingga kurva PR-nya lebih penuh. *fear* berada di tengah (F1 0.459; AUC 0.847), sedangkan *anger* 0.418 dan *anticipation* 0.408 mengisyaratkan ambiguitas leksikal atau ketergantungan konteks yang lebih kuat.

Dua label paling sulit adalah *surprise* (F1 0.118; AUC 0.855) dan *disgust* (gagal terdeteksi pada test, F1 ≈ 0 karena sangat jarang). *Surprise* kerap muncul dalam ragam tuturan yang tidak selalu memakai kata “kaget/terkejut”, sehingga sinyalnya tipis. Sementara itu, rendahnya prevalensi *disgust* membuat model tidak menemukan kasus positif pada *test*, sehingga F1 menjadi nol dan AUC berpotensi tidak stabil. Kondisi ini memotivasi penerapan kalibrasi khusus pasca pelatihan di luar penyesuaian pelatihan murni.

3.3 Analisis Per label

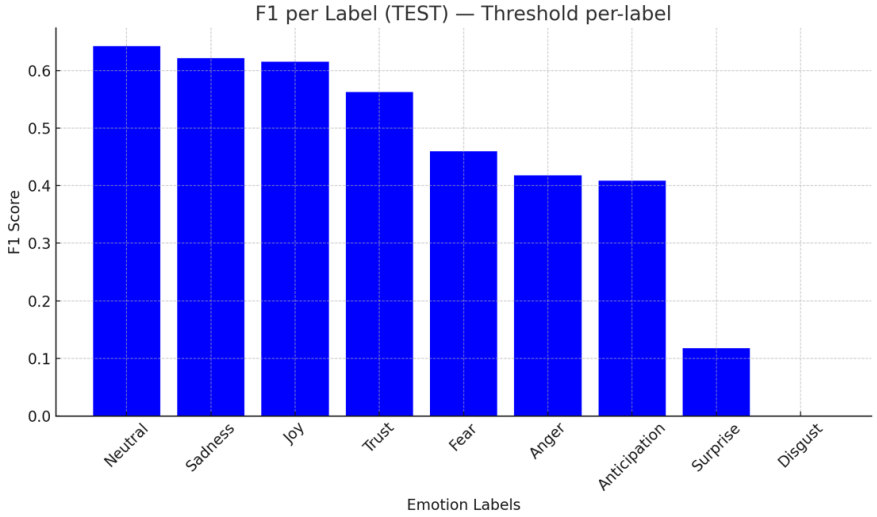
Performa antar label bervariasi. *Neutral* (F1 0.642), *sadness* (0.622), *joy* (0.615), dan *trust* (0.563) tampil kuat dengan ROC-AUC tinggi. *Fear* (0.459) berada di tingkat menengah, sedangkan *anger* (0.418) dan *anticipation* (0.408) cenderung lebih menantang. *Surprise* (0.118) sulit karena sinyal leksikal yang tipis atau kontekstual, sementara *disgust* nyaris tidak muncul di *test* sehingga F1 awal tidak informatif.

Pola ini selaras dengan profil IAA: kategori dengan κ tinggi (*neutral*, *sadness*, *joy*) cenderung menunjukkan F1 yang lebih stabil, sedangkan kategori dengan κ *moderate* (*surprise, anticipation, disgust*) memerlukan kalibrasi per label untuk memperbaiki *operating point*. *Trust* dan *fear* berada di tingkat menengah dengan AUC tinggi namun F1 tidak setinggi kelompok κ tinggi, sejalan dengan temuan GoEmotions tentang stabilitas emosi eksplisit [23] dan praktik per-label *thresholding* pada pembelajaran multi-label [22].

Secara per-label, beberapa emosi tampil menonjol: *neutral* memperoleh F1 0.642 (*precision* 0.599; *recall* 0.693; AUC 0.834), *sadness* 0.622 (0.531; 0.750; AUC 0.906), *joy* 0.615 (0.543; 0.710; AUC 0.911), dan *trust* 0.563 (0.493; 0.655; AUC 0.905). Keempatnya umumnya hadir dengan penanda kata yang eksplisit atau pola pragmatik yang relatif konsisten, sehingga kurva PR-nya lebih penuh. *Fear* berada di tengah (F1 0.459; AUC 0.847), sedangkan *anger* 0.418 dan *anticipation* 0.408 mengisyaratkan ambiguitas leksikal atau ketergantungan konteks yang lebih kuat.

Dua label paling sulit adalah *surprise* (F1 0.118; AUC 0.855) dan *disgust* (gagal terdeteksi pada test, F1 ≈ 0 karena sangat jarang). *Surprise* kerap muncul dalam ragam tuturan yang tidak selalu memakai kata "kaget/terkejut", sehingga sinyalnya tipis. Sementara itu, rendahnya prevalensi *disgust* membuat model tidak

menemukan kasus positif pada test, sehingga F1 menjadi nol dan AUC berpotensi tidak stabil. Kondisi ini memotivasi penerapan kalibrasi khusus pasca pelatihan di luar penyesuaian pelatihan murni.



Gambar 4. Skor F1 per label pada set uji (ambang per-label)

Tabel 6 Matriks hasil per label pada *test set* (n=300)

Label	*Precision*	*Recall*	F1	*Support*	ROC-AUC	Kappa (IAA, N=600)
Neutral	0.5988	0.6929	0.6424	194	0.8337	0.80
Sadness	0.5310	0.7500	0.6218	19	0.9062	0.73
Joy	0.5432	0.7097	0.6154	34	0.9110	0.70
Trust	0.4932	0.6545	0.5625	16	0.9049	0.62
Fear	0.5152	0.4146	0.4595	10	0.8467	0.59
Anger	0.4043	0.4318	0.4176	12	0.7835	0.66
Anticipation	0.3537	0.4833	0.4085	16	0.7692	0.56
Surprise	0.4000	0.0690	0.1176	32	0.8553	0.53
Disgust	0.0000	0.0000	0.0000	5	0.6855	0.46

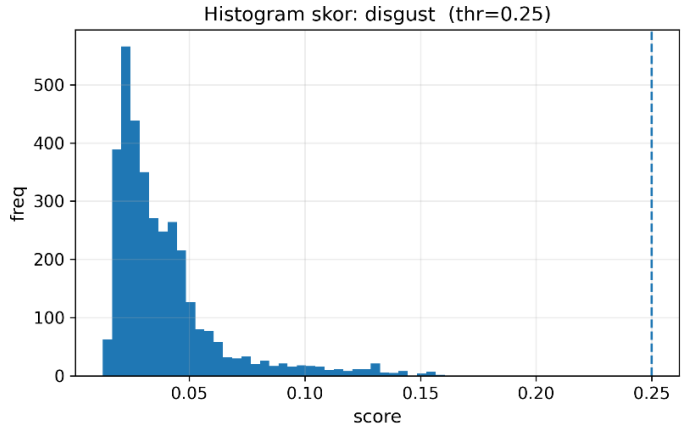
Support menunjukkan jumlah sampel positif per label di test set (n=300). Total support = 338 karena multi-label.

3.4 Penanganan Label Minoritas (*Disgust*)

Prevalensi yang sangat rendah pada *disgust* membuat metrik berbasis keseimbangan, terutama F1 dan kurva *precision-recall*, kurang sensitif: perubahan kecil pada jumlah observasi dapat memicu fluktuasi tajam dan menurunkan *recall* ketika memakai ambang global, meski separabilitas skor masih terindikasi [22][23]. Pada set *dev*, distribusi skor *disgust* terpusat di bawah 0.10. Dengan ambang 0.25, *recall* praktis nol.

Untuk mengatasinya, kami menerapkan kalibrasi bertingkat: (i) *post-hoc calibration* per label menggunakan *Platt scaling* dengan target presisi tinggi guna menahan *false positive* [9],[10]; (ii) *rate-targeting* pada korpus aplikasi melalui pemilihan ambang berbasis kuantil skor agar prevalensi prediksi selaras *base rate* empiris pada data target (target ~2-3%) [6],[24]; dan (iii) *lexicon-aware boost* berbasis kata kunci (mis. muak, mual, jijik, najis, eww) dengan pembatasan konteks, diikuti final clamp pada probabilitas atau label sebagai heuristik pasca proses yang ringan dan model-*agnostic* [6],[24].

Penetapan ambang akhir = 0.122 memulihkan *recall* tanpa melambungkan kesalahan positif, sekaligus menjaga konsistensi agregat lintas label. Praktik *per-label thresholding* untuk kelas langka ini sejalan dengan literatur multi-label dan studi emosi berbutir halus [22],[23],[24].

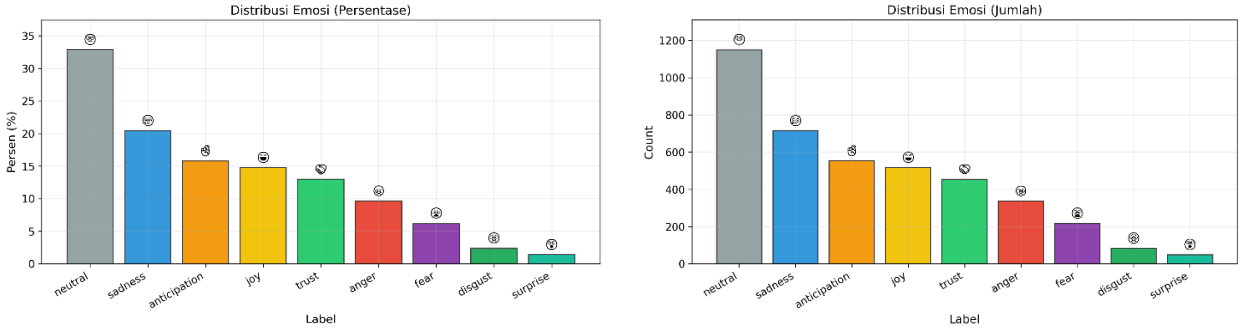


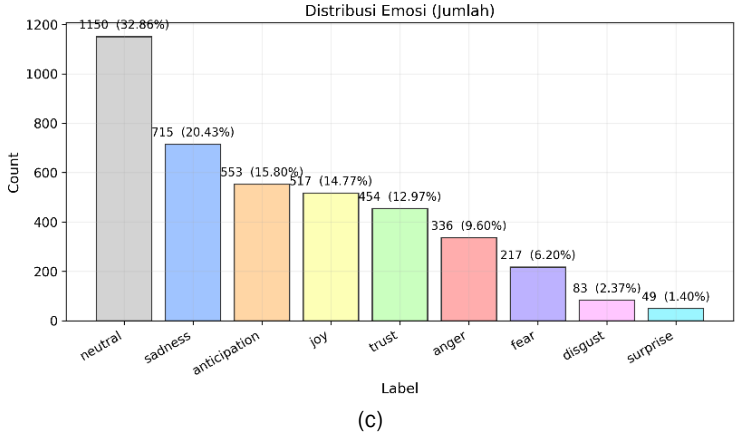
Gambar 5. Histogram skor *disgust* (*dev*) dengan penanda ambang 0.25.

3.5 Hasil pada Korpus Aplikasi 3500 Data

Pada korpus aplikasi gabungan (N = 3500; terdiri atas 600 data beranotasi manual dan 2900 data aplikasi), setelah kalibrasi per label, penetapan ambang akhir untuk *disgust* (= 0.122), serta penegakan aturan *neutral-exclusive* dan anti-kosong, diperoleh prevalensi: *neutral* 32.86%, *sadness* 20.43%, *anticipation* 15.80%, *joy* 14.77%, *trust* 12.97%, *anger* 9.60%, *fear* 6.20%, *surprise* 1.40%, dan *disgust* 2.37%.

Angka-angka ini menyatakan proporsi dokumen yang memuat tiap label dalam skema multi-label, sehingga total persentase dapat melampaui 100%. Pola ini sejalan dengan karakter domain: tuturan informatif atau bernada netral dominan, diikuti emosi sedih dan antisipatif; disgust tetap jarang namun terkendali (83 dari 3500; 2.37%).





(c)

Gambar 6. Visualisasi data: (a) persentase distribusi emosi, (b) jumlah distribusi emosi, (c) persentase dan jumlah distribusi emosi

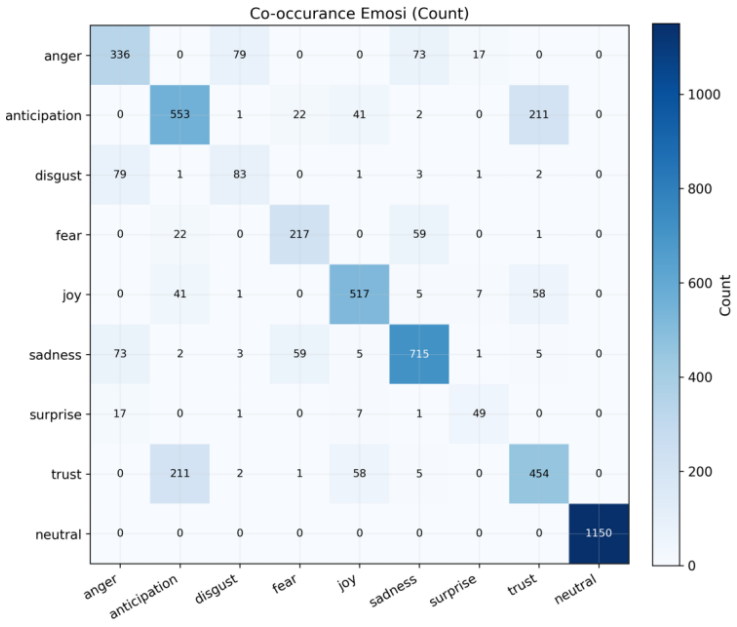
Tabel 7. Hasil persentase dan jumlah distribusi emosi

No	Label	Jumlah	Persentase
1	*Neutral*	1150	32.86
2	*Sadness*	715	20.43
3	*Anticipation*	553	15.8
4	*Joy*	517	14.77
5	*Trust*	454	12.97
6	*Anger*	336	9.6
7	*Fear*	217	6.2
8	*Disgust*	83	2.37
9	*Surprise*	49	1.4

3.6 *Co-occurrence* Antar Emosi

Analisis *co-occurrence* antar emosi pada korpus aplikasi menunjukkan pola pasangan yang sejalan dengan intuisi domain, misalnya *joy-trust* (kebahagian disertai keyakinan) dan *sadness-anger* (kesedihan atau kemarahan). Ketergantungan antar label semacam ini merupakan sinyal penting dalam multi-label *classification* (MLC) dan kerap dimodelkan untuk meningkatkan kinerja prediksi [1],[25].

Perlu dicatat, *neutral-exclusive* membatasi *co-occurrence* neutral dengan label lain sehingga sel terkait neutral pada matriks co-occurrence didominasi *diagonal, sementara rendahnya co-occurrence disgust konsisten dengan prevalensinya yang jarang.*



Gambar 7. Peta *heatmap co-occurrence* emosi (jumlah dokumen, N=3,500).

3.7 Pembahasan

3.7.1 Pencapaian Tujuan Penelitian

Penelitian ini bertujuan menghasilkan pemodelan emosi yang reliabel pada korpus media sosial *long-tail* serta menunjukkan praktik kalibrasi yang dapat digeneralisasi untuk bahasa Indonesia. Tujuan pertama tercapai melalui perolehan *micro-F1* = 0.542 dan *macro-F1* = 0.427 pada test set, dengan *precision micro* = 0.510, *recall micro* = 0.579, dan *Hamming loss* = 0.109. Nilai *micro-F1* yang lebih tinggi dibanding *macro-F1* menunjukkan model lebih efektif pada kelas mayoritas, sedangkan *macro-F1* mengonfirmasi masih adanya disparitas pada kelas minoritas, sesuai karakteristik data *long-tail* [18].

Tujuan kedua, yaitu menunjukkan praktik kalibrasi yang dapat digeneralisasi, terwujud melalui strategi pasca pelatihan yang memadukan kalibrasi posterior (*Platt/Isotonic*), per-label *thresholding* bertarget presisi, *rate-targeting* berbasis kuantil skor, serta *lexicon-aware boost* terbatas konteks. Strategi ini bersifat *model-agnostic*

karena tidak mengubah arsitektur atau bobot model, sehingga dapat diterapkan pada model dasar lain atau korpus baru yang mengalami *distribution shift* [6],[14],[15].

3.7.2 Solusi terhadap Masalah Penelitian

Masalah 1: Ketidakseimbangan label ekstrem pada data *long-tail*

Masalah ini dihadapi melalui tiga strategi. Pertama, penyetelan ambang per-label berbasis *development set* menghasilkan ambang optimal yang bervariasi: umumnya 0.25, dengan 0.23 untuk *surprise* dan 0.122 untuk *disgust* setelah kalibrasi bertarget. Kurva F1 terhadap *threshold* menunjukkan *plateau* stabil di kisaran 0.20–0.30 (Gambar 2), mengonfirmasi efektivitas strategi ini dalam menyeimbangkan *precision* dan *recall* lintas label.

Kedua, untuk label paling jarang (*disgust*), diterapkan kalibrasi bertarget yang menggabungkan *precision-target* (≥ 0.80) di *dev*, *rate-target* pada korpus aplikasi (target 2.5% ± 4%), *lexicon boost* berbasis kata kunci (muak, mual, jijik, najis, *eww*) dengan pembatasan konteks, dan *final clamp*. Hasilnya, prevalensi *disgust* pada 3500 data aplikasi stabil di 2.37% (83 dokumen), mendekati target 2.5% tanpa lonjakan *false positive*. Histogram skor *disgust* pada *dev* (Gambar 5) menunjukkan distribusi terpusat di bawah 0.10, sehingga ambang global 0.25 akan menghasilkan *recall* mendekati nol—kondisi yang berhasil diatasi melalui kalibrasi khusus.

Ketiga, aturan *neutral-exclusive* (neutral aktif hanya ketika emosi lain tidak melewati ambang) dan anti-kosong (mengaktifkan neutral ketika seluruh skor di bawah ambang) memastikan setiap dokumen memiliki minimal satu label, mencegah keluaran kosong yang kerap terjadi pada data *in-the-wild*.

Masalah 2: *Distribution shift* antara korpus pelatihan dan aplikasi

Fenomena *label shift* atau *prevalence shift* ditangani melalui *rate-targeting* berbasis kuantil skor yang menyelaraskan distribusi prediksi dengan *base rate* empiris korpus aplikasi tanpa memerlukan label target [7],[8]. Pendekatan ini lebih ringan dibanding teknik *quantification* yang memerlukan estimasi kompleks, namun tetap efektif dalam menjaga proporsi kelas minoritas.

Masalah 3: Sinyal leksikal tipis untuk emosi tertentu

Emosi seperti *surprise* (F1 = 0.118) dan *anticipation* (F1 = 0.408) menunjukkan performa lebih rendah karena tidak selalu memakai penanda kata eksplisit. Sebaliknya, *neutral* (F1 = 0.642), *sadness* (F1 = 0.622), *joy* (F1 = 0.615), dan *trust* (F1 = 0.563) mencapai F1 ≥ 0.56 dengan ROC-AUC tinggi (0.834–0.911), menandakan pola yang lebih konsisten dan penanda leksikal yang lebih jelas. Pola "mudah vs. samar" ini sejalan dengan temuan di GoEmotions bahwa emosi eksplisit cenderung lebih stabil ketimbang kategori kontekstual [9],[20].

3.7.3 Perbandingan dengan Riset Sebelumnya

Temuan penelitian ini konsisten dengan studi emosi multi-label mutakhir dalam tiga aspek. Pertama, efektivitas *transfer learning* berbasis Transformer untuk data media sosial dan bahasa non-Inggris [12],[13], [21], [25]. Pemilihan IndoBERTweet sebagai *backbone* selaras dengan domain cuitan dan ekosistem evaluasi IndoNLU [2],[3], dengan keunggulan mengatasi *vocabulary mismatch* melalui inisialisasi kosakata domain spesifik.

Kedua, pentingnya kalibrasi pasca pelatihan disertai *per-label thresholding* pada distribusi *long-tail* [6],[23], [24]. Penelitian ini memperluas praktik tersebut dengan menggabungkan *rate-targeting* dan *lexicon boost* terbatas konteks, menghasilkan *pipeline* yang lebih komprehensif untuk kelas minoritas.

Ketiga, relevansi analisis *co-occurrence* dalam MLC [1],[25]. Heatmap *co-occurrence* (Gambar 7) mengonfirmasi pasangan emosi yang sejalan dengan intuisi domain (*joy–trust*, *sadness–anger*), memberikan validasi eksternal terhadap kerangka Plutchik dan menunjukkan model mampu menangkap ketergantungan antar label meskipun dilatih secara independen (*BCEWithLogitsLoss* tidak memodelkan korelasi eksplisit).

Namun, penelitian ini berkontribusi pada aspek yang kurang dibahas literatur sebelumnya, yaitu kombinasi sistematis kalibrasi posterior, *threshold tuning*, *rate-targeting*, dan *lexicon boost* dalam satu *pipeline* pasca pelatihan yang ringan dan *model-agnostic* untuk bahasa Indonesia. Kebanyakan studi fokus pada satu atau dua teknik saja, sedangkan pendekatan holistik ini memungkinkan penanganan simultan terhadap *class imbalance*, *distribution shift*, dan *weak lexical signal*.

3.7.4 Implikasi Praktis dan Teoritis

Implikasi praktis: Pemetaan emosi pada 3500 cuitan menghasilkan distribusi yang masuk akal untuk konteks pengumuman UTBK: *neutral* dominan (32.86%), diikuti emosi negatif seperti *sadness* (20.43%) dan *anger* (9.60%), serta emosi positif seperti *joy* (14.77%) dan *trust* (12.97%). Prevalensi *anticipation* yang cukup tinggi (15.80%) mencerminkan dinamika temporal di sekitar pengumuman—kombinasi harapan, kecemasan, dan persiapan untuk tahap berikutnya. Pola ini memberikan *insight* bagi pemangku kepentingan pendidikan tinggi untuk memahami respons emosional siswa dan merancang komunikasi yang lebih empatik di sekitar periode pengumuman.

Implikasi teoritis: Konsistensi pola *co-occurrence* dengan roda emosi Plutchik (Gambar 7) memberikan validasi eksternal terhadap kerangka teoretis, meskipun model tidak secara eksplisit memodelkan korelasi antar label. Hal ini menunjukkan bahwa representasi kontekstual dari IndoBERTweet sudah menangkap ketergantungan emosi pada tingkat semantik, sejalan dengan temuan bahwa *pretrained embeddings* secara implisit mengkodekan struktur afektif bahasa [12], [20].

Keberhasilan kalibrasi pasca pelatihan dalam meningkatkan deteksi kelas minoritas tanpa *fine-tuning* ulang memperkuat argumen bahwa *operating point selection* sama pentingnya dengan arsitektur model dalam klasifikasi multi-label [6], [22]. Temuan ini relevan untuk praktisi yang bekerja dengan sumber daya komputasi terbatas atau tidak memiliki akses untuk melatih ulang model besar.

3.8 Keterbatasan dan Peluang Pengembangan

Batasan utama penelitian ini adalah ketidakseimbangan label pada himpunan pengembangan dan uji, khususnya rendahnya prevalensi *disgust* yang menjadikan metrik F1 kurang sensitif. Untuk memitigasi hal tersebut, interpretasi hasil dilengkapi dengan kurva *precision-recall* (PR), prosedur kalibrasi pasca pelatihan, dan evaluasi pada korpus aplikasi [6], [9], [10].

Keterbatasan lain meliputi *scope* penelitian yang fokus pada emosi eksplisit sesuai kerangka Plutchik (8 emosi dasar + *neutral*), sehingga fenomena linguistik kompleks seperti sarkasme, ironi, dan negasi tidak menjadi target deteksi dalam penelitian ini. Anotator tidak ditugaskan menandai sarkasme karena bukan merupakan kategori emosi dalam taksonomi yang digunakan. Pengujian juga dilakukan pada satu konfigurasi inti (*seed / hiperparameter*) dan satu domain aplikasi. Ke depan, peningkatan reliabilitas dapat diupayakan melalui: (i) perluasan *gold subset* terutama untuk *disgust* dan *surprise*; (ii) eksplorasi ekstensi label untuk menangkap fenomena linguistik kompleks (sarkasme, ironi) sebagai kategori terpisah, yang memerlukan pedoman anotasi khusus dan *epoch* pelatihan lebih banyak; (iii) penambahan modul *preprocessing* untuk menangani negasi; (iv) uji *robustness* terhadap variasi *seed / hiperparameter* dan lintas domain; serta (v) eksplorasi model yang memanfaatkan korelasi antar-label secara eksplisit. Arah ini sejalan dengan rekomendasi terkini dalam *multi-label classification* dan kalibrasi model [6],[10],[12],[13], [21]–[23], [25].

4 Kesimpulan

Penelitian ini berhasil menjawab tiga masalah penelitian yang dirumuskan:

Kesimpulan 1 (Masalah 1 & Tujuan 1): Strategi kalibrasi pasca pelatihan bertarget yang mengkombinasikan *precision-target, rate-target, dan lexicon-aware boost* efektif menangani ketidakseimbangan label ekstrem pada data *long-tail*. Untuk label minoritas *disgust* (prevalensi 2.2%), ambang khusus 0.122 berhasil memulihkan *recall* tanpa lonjakan *false positive*, menghasilkan prevalensi prediksi 2.37% yang mendekati *base rate* tanpa pelatihan ulang. Model mencapai keseimbangan kinerja antara metrik mikro (*micro-F1* = 0.542) dan makro (*macro-F1* = 0.427).

Kesimpulan 2 (Masalah 2 & Tujuan 2): *Rate-targeting* berbasis kuantil skor terbukti dapat mengatasi distribution shift antara korpus pelatihan dan aplikasi tanpa memerlukan label target. *Pipeline* bersifat *model-agnostic* dan ringan (tidak mengubah bobot model), sehingga dapat diterapkan pada korpus baru dengan karakteristik distribusi berbeda. Aturan *neutral-exclusive* dan mekanisme anti-kosong membuat keputusan label lebih konsisten pada skenario dunia nyata.

Kesimpulan 3 (Masalah 3 & Tujuan 3): Model IndoBERTweet dengan kalibrasi pasca pelatihan mampu menangkap pola emosi yang konsisten dengan kerangka teoretis Plutchik. Analisis *co-occurrence* mengonfirmasi pasangan emosi intuitif (*joy-trust, sadness-anger*), meskipun emosi dengan sinyal leksikal tipis (surprise F1=0.118, *anticipation* F1=0.408) masih memerlukan strategi tambahan. Pemetaan emosi pada 3500 cuitan menunjukkan profil distribusi yang konsisten dengan karakteristik korpus aplikasi: *neutral* dominan (32.86%), diikuti *sadness* (20.43%), *anticipation* (15.80%), *joy* (14.77%), dan *trust* (12.97%).

Arah pengembangan ke depan mencakup perluasan *subset* anotasi manual untuk kelas langka (*disgust*, *surprise*), eksplorasi ekstensi label untuk menangkap fenomena linguistik kompleks (sarkasme, ironi) sebagai kategori terpisah yang memerlukan pedoman anotasi khusus dan *epoch* pelatihan lebih banyak, penambahan modul *preprocessing* untuk menangani negasi, uji *robustness* lintas domain dan variasi *seed* / *hiperparameter*, serta eksplorasi pemodelan dependensi antar-label secara eksplisit (misalnya melalui *Classifier Chains* atau *attention mechanism*).

Untuk mendukung replikasi, penulis menyediakan paket rilis yang berisi model terkalibrasi (*checkpoint-616*), konfigurasi pasca proses (ambang per label, pola leksikon, aturan *neutral-exclusive*), dan skrip inferensi sehingga seluruh langkah dapat diulang dengan pengaturan yang sama pada korpus baru.

Referensi

[1] A. Semeraro, S. Vilella, and G. Ruffo, "PyPlutchik: Visualising and comparing emotion-annotated corpora," *PLOS ONE*, vol. 16, no. 9, p. e0256503, Sep. 2021. https://doi.org/10.1371/journal.pone.0256503

[2] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pp. 10660–10668, Nov. 7–11, 2021. [Online] Available : https://aclanthology.org/2021.emnlp-main.833/

[3] S. Cahyawijaya, H. Lovenia, A. F. Aji, G. I. Winata, B. Wilie, F. Koto, R. Mahendra, et al., "NusaCrowd: open source initiative for Indonesian NLP resources," *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, pp. 13745–13818, 2023. https://doi.org/10.18653/v1/2023.findings-acl.868

[4] P. W. Koh, S. Sagawa, H. Marklund, M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al., "WILDS: A benchmark of in-the-wild distribution shifts," in *Proc. 38th Int. Conf. Machine Learning (ICML 2021)*, PMLR 139, pp. 5637–5664, 2021. [Online] Available : https://proceedings.mlr.press/v139/koh21a/koh21a.pdf

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pp. 4171–4186, 2019. https://doi.org/10.48550/arXiv.1810.04805

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *in Proc. 34th Int. Conf. Machine Learning (ICML 2017),* PMLR 70, Sydney, Australia, 2017. [Online] Available : https://proceedings.mlr.press/v70/guo17a/guo17a.pdf

[7] K. Draszawka and J. Szymański, "From scores to predictions in multi-label classification: neural thresholding strategies," *Applied Sciences*, vol. 13, no. 13, p. 7591, 2023. https://doi.org/10.3390/app13137591

[8] S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton, "A unified view of label shift estimation," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020. [Online] Available : https://proceedings.neurips.cc/paper/2020/

[9] K. Draszawka and J. Szymański, "From scores to predictions in multi-label classification: neural thresholding strategies," *Applied Sciences*, vol. 13, no. 13, p. 7591, 2023. https://doi.org/10.3390/app13137591

[10] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLOS ONE*, vol. 17, no. 1, p. e0262838, 2022. https://doi.org/10.1371/journal.pone.0262838

[11] R. Riccosan and K. E. Saputra, "Multilabel multiclass sentiment and emotion dataset from Indonesian mobile application review," *Data in Brief*, vol. 50, Art. no. 109576, 2023, https://doi.org/10.1016/j.dib.2023.109576

[12] A. K. Glenn, P. LaCasse, and B. Cox, "Emotion classification of Indonesian tweets using bidirectional LSTM," *Neural Computing and Applications*, vol. 35, pp. 9567–9578, 2023. https://doi.org/10.1007/s00521-022-08186-1

[13] M. H. Algifari and E. D. Nugroho, "Emotion classification of Indonesian tweets using BERT embedding," *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 172–176, 2023. https://doi.org/10.30871/jaic.v7i2.6528

[14] Z. C. Lipton, Y.-X. Wang, and A. J. Smola, "Detecting and correcting for label shift with Black Box Predictors," in *Proc. 35th Int. Conf. Machine Learning (ICML 2018)*, PMLR 80, Stockholm, Sweden, 2018, [Online] Available : https://proceedings.mlr.press/v80/lipton18a/lipton18a.pdf

[15] C. Wang, "Calibration in Deep Learning: a survey of the state-of-the-art," Arxiv: 2308.01222, 2025. https://doi.org/10.48550/arXiv.2308.01222

[16]J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (*4th ed.). Sage Publications, 2014. [Online] Available : https://www.ucg.ac.me/skladiste
[17]JustAnotherArchivist, "Snscrape: A social networking service scraper in Python," *GitHub repository*, 2021. [Online]. Available: https://github.com/JustAnotherArchivist/snscrape. Accessed: Sep. 26, 2025.
[18]S. Kishore, D. Sundaram, and M. D. Myers, "A temporal dynamics framework and methodology for computationally intensive social media research," *Journal of Information Technology*, vol. 40, no. 2, pp. 140–163, 2025. https://doi.org/10.1177/02683962241283051
[19]L. Bozarth and C. Budak, "Keyword expansion techniques for collecting Twitter data on social movements," *EPJ Data Science*, vol. 11, no. 30, 2022. https://doi.org/10.1140/epjds/s13688-022-00343-9
[20]Paula Vicente, "Sampling Twitter users for social science research: evidence from a systematic review of the literature," *Quality & Quantity*, vol. 57, no. 6, pp. 5449–5489, 2023. https://doi.org/10.1007/s11135-023-01615-w
[21] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013. https://doi.org/10.1111/j.1467-8640.2012.00460.x
[22] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Social Network Analysis and Mining*, vol. 13, no. 31, 2023. https://doi.org/10.1007/s13278-023-01030-x
[23] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted F1-scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, Oct. 2024. https://doi.org/10.3390/app14219863
[24] B. Wilie *et al*., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 1st Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and 10th Int. Joint Conf. on Natural Language Processing (AACL-IJCNLP)*, Suzhou, China, pp. 843–857, 2020. [Online] Available : https://aclanthology.org/2020.aacl-main.85/
[25] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: a dataset of fine-grained emotions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL 2020)*,pp. 4040–4054 , Jul. 2020. https://doi.org/10.18653/v1/2020.acl-main.372