Jurnal Publikasi Ilmu Komputer dan Multimedia

E-ISSN: 2808-8999
P-ISSN: 2808-9375

Analisis Sentimen dan Emosi Publik terhadap Pelaksanaan Program Makan Bergizi Gratis pada Media Sosial Threads

Harianto[1] , Virda Mega Ayu[2], Muchammad Hamdani[3], Mujito[4]

[1234]ITB Dewantara, Acropolis, RUKO Blk. LC, Jl. Raya Karadenan Jl. Bojong Depok Baru III Blok LC 19, Karadenan, Kec. Cibinong, Kabupaten Bogor, Jawa Barat 16913
* Penulis : Harianto

Abstract:
Indonesia's Free Nutritious Meal Program (Makan Bergizi Gratis, MBG) is a national strategic policy targeting 82.9 million beneficiaries by 2026, and its scale generates substantial public discussion on social media. Objective. This study characterises public perception of MBG implementation on Threads through relevance detection, sentiment classification and emotion classification of Indonesian-language posts. Methods. Following the CRISP-DM framework, 7,016 Threads posts were retrieved on 20 June 2026 using eight MBG-related keywords. Labels for relevance, sentiment (positive, negative, neutral, mixed) and emotion were produced by an LLM-assisted annotation pipeline under a fixed schema and subsequently verified by the authors on a stratified sample. After excluding one post with a missing relevance label and 165 posts left empty by text cleaning, 6,850 posts entered the relevance task and the 4,968 relevant, non-empty posts entered the sentiment and emotion tasks. Five classifiers — Logistic Regression, Linear SVC, Multinomial Naive Bayes, Complement Naive Bayes and Random Forest — were compared on TF-IDF features (10,000 terms, unigram + bigram). Evaluation used Macro F1, Weighted F1, per-class metrics and 5-fold cross-validation on the training partition, with a single held-out test evaluation. A chi-square test of independence with Cramér's V quantified the sentiment–emotion association. Results. Model selection was performed on the validation partition, where Random Forest led relevance detection (Macro F1 = 0.8751) and Logistic Regression led sentiment (0.5206) and emotion (0.2048). On the held-out test partition, evaluated once, the selected models reached Macro F1 = 0.8904 for relevance (n = 1,028), 0.4810 for sentiment and 0.2382 for emotion (n = 746 each); the low emotion score reflects 15 highly imbalanced classes and is reported as an exploratory result. Within the 4,968 relevant posts, negative sentiment accounted for 54.3% (2,696), and frustration (37.7%; 1,873) and anger (11.4%; 565) were the most frequent emotions; the four risk-indicating emotions (frustration, anger, disappointment, worry) together accounted for 57.1% (2,836). Sentiment and emotion were strongly associated (χ^2 = 8,313.14, df = 81, N = 4,968, $p < 0.001$, Cramér's V = 0.747). Terms most strongly associated with the negative class included "korupsi", "proyek", "keracunan" and "anggaran". Conclusion. Within this Threads dataset, criticism-oriented discourse dominates the MBG conversation, and emotion labels add discriminative information beyond three-way sentiment. Findings are descriptive and platform-specific, and are offered as one indicative input among others for policy communication rather than as evidence of causal effects.

Keywords: sentiment analysis; emotion classification; CRISP-DM; free nutritious meal program; public policy; Threads; Indonesian NLP

Diterima: 21 Juni 2026
Direvisi: 11 Agustus 2026
Diterima: 2 September 2026
Diterbitkan: 7 September 2026
Versi sekarang: 30 Sept 2026

Abstrak:
Program Makan Bergizi Gratis (MBG) merupakan kebijakan strategis nasional yang menargetkan 82,9 juta penerima manfaat pada 2026, dan skalanya memunculkan diskusi publik yang luas di media sosial. Tujuan. Penelitian ini memetakan persepsi publik terhadap pelaksanaan MBG di Threads melalui deteksi relevansi, klasifikasi sentimen, dan klasifikasi emosi pada unggahan berbahasa Indonesia. Metode. Mengikuti kerangka CRISP-DM, 7.016 unggahan Threads ditarik pada 20 Juni 2026 menggunakan delapan kata kunci terkait MBG. Label relevansi, sentimen (positive, negative, neutral, mixed), dan emosi dihasilkan melalui pipeline anotasi berbantuan LLM dengan skema tetap, lalu diverifikasi penulis pada sampel terstratifikasi. Setelah mengeluarkan 1 unggahan tanpa label relevansi dan 165 unggahan yang menjadi kosong setelah pembersihan teks, 6.850 unggahan masuk ke tugas relevansi dan 4.968 unggahan relevan non-kosong masuk ke tugas sentimen dan emosi. Lima pengklasifikasi — Logistic Regression, Linear SVC, Multinomial Naive Bayes, Complement Naive Bayes, dan Random Forest — dibandingkan pada fitur TF-IDF (10.000 term, unigram + bigram). Evaluasi menggunakan Macro F1, Weighted F1, metrik per kelas, dan 5-fold cross-validation pada partisi latih, dengan satu kali evaluasi pada data uji yang ditahan. Uji khi-kuadrat independensi beserta Cramér's V digunakan untuk mengukur asosiasi sentimen–emosi. Hasil. Pemilihan model dilakukan pada partisi validasi, dengan Random Forest unggul pada deteksi relevansi (Macro F1 = 0,8751) serta Logistic Regression unggul pada sentimen (0,5206) dan emosi (0,2048). Pada partisi uji yang ditahan dan dievaluasi satu kali, model terpilih mencapai Macro F1 = 0,8904 untuk relevansi (n = 1.028), 0,4810 untuk sentimen, dan 0,2382 untuk emosi (n = 746 untuk keduanya); skor emosi yang rendah mencerminkan 15 kelas yang sangat tidak seimbang dan dilaporkan sebagai hasil eksploratif. Pada 4.968 unggahan relevan, sentimen negatif mencapai 54,3% (2.696), dengan frustration 37,7% (1.873) dan anger 11,4% (565) sebagai emosi terbanyak; gabungan empat emosi berisiko (frustration, anger, disappointment, worry) mencapai 57,1% (2.836). Sentimen dan emosi berasosiasi kuat (χ^2 = 8.313,14; df = 81; N = 4.968; $p < 0,001$; Cramér's V = 0,747). Term yang paling berasosiasi dengan kelas negatif antara lain "korupsi", "proyek", "keracunan", dan "anggaran". Kesimpulan. Dalam dataset Threads ini, wacana bernada kritik mendominasi percakapan MBG, dan label emosi menambah informasi pembeda di luar sentimen tiga arah. Temuan bersifat deskriptif dan spesifik platform, serta ditawarkan sebagai satu masukan indikatif bagi komunikasi kebijakan, bukan sebagai bukti hubungan kausal.
Kata kunci: analisis sentimen; klasifikasi emosi; CRISP-DM; program MBG; kebijakan publik; Threads; NLP Indonesia

1. Pendahuluan

Program Makan Bergizi Gratis (MBG) merupakan salah satu kebijakan strategis nasional Indonesia yang bertujuan meningkatkan kualitas sumber daya manusia melalui pemenuhan gizi masyarakat, khususnya bagi kelompok penerima manfaat prioritas. Pada tahun 2026, Badan Gizi Nasional menargetkan program ini menjangkau 82,9 juta penerima manfaat [1], setelah pada awal tahun yang sama program tersebut dilaporkan menjangkau hampir 60 juta penerima manfaat [18]. Dengan skala implementasi sebesar itu, MBG bukan hanya program bantuan pangan, melainkan kebijakan publik dengan dampak sosial, ekonomi, kesehatan, dan politik yang luas.

Dalam implementasinya, program MBG menghadapi berbagai tantangan, seperti tata kelola distribusi, validasi data penerima, pengawasan mutu pangan, serta keamanan konsumsi. Pemerintah terus mendorong penguatan pengawasan program untuk memastikan kualitas layanan tetap terjaga [2]. Namun demikian, sejumlah kasus dugaan keracunan pangan dan berbagai kritik terhadap implementasi program memunculkan respons publik yang beragam [3], dan evaluasi pemerintah sendiri menyoroti persoalan data penerima serta mutu layanan [20]. Kondisi ini menunjukkan bahwa evaluasi kebijakan MBG perlu dilakukan tidak hanya dari sisi operasional, tetapi juga dari sisi persepsi publik.

Media sosial menjadi salah satu sumber data untuk memahami persepsi masyarakat terhadap kebijakan publik. Berbeda dengan survei konvensional, media sosial memungkinkan pengambilan opini secara lebih luas, spontan, dan mendekati waktu nyata, meskipun dengan konsekuensi bias representasi yang harus diperhitungkan. Salah satu platform yang berkembang pesat adalah Threads, yaitu platform mikroblog berbasis teks dari Meta yang banyak digunakan untuk diskusi sosial secara naratif dan personal [14][15]. Karakteristik naratif tersebut membuat unggahan Threads cenderung lebih panjang dan lebih argumentatif dibandingkan unggahan pada platform mikroblog lain, sehingga relevan sebagai sumber data analisis persepsi publik.

1.1. Kesenjangan penelitian dan kebaruan

Penelitian analisis sentimen terhadap program MBG telah dilakukan pada beberapa platform. Nugroho dkk. [4] mengklasifikasikan opini publik pada komentar YouTube menggunakan LSTM; Putri dkk. [5] membandingkan pelabelan leksikon InSet dan VADER dengan SVM pada platform X; Muhabbab dkk. [6] dan Ananda dkk. [7] menerapkan IndoBERT dan NusaBERT pada data X; Pratama dkk. [8] menggunakan BERTopic untuk pemodelan topik pada 19.843 komentar YouTube; dan Kusuma dkk. [9] menganalisis respons publik pada data Twitter. Tabel 1 merangkum posisi penelitian-penelitian tersebut.

Tabel 1. Sintesis penelitian terdahulu tentang persepsi publik terhadap MBG dan posisi penelitian ini

Studi	Platform	Ukuran data	Jenis label	Model	Keluaran utama
Nugroho dkk. [4]	YouTube	7.733 komentar	Sentimen 3 kelas	LSTM	Klasifikasi sentimen
Putri dkk. [5]	X/Twitter	±1.000–3.000 tweet	Sentimen leksikon (InSet, VADER)	SVM	Perbandingan skema pelabelan
Muhabbab dkk. [6]	X/Twitter	±5.000 tweet	Sentimen 3 kelas	IndoBERT	Peningkatan akurasi transformer
Ananda dkk. [7]	X/Twitter	±10.000 tweet	Sentimen 3 kelas	IndoBERT-Large, NusaBERT-Large	Perbandingan model besar
Pratama dkk. [8]	YouTube	19.843 komentar	Topik (tanpa penyelia)	BERTopic	Aspek/topik implementasi

Studi	Platform	Ukuran data	Jenis label	Model	Keluaran utama
Kusuma dkk. [9]	X/Twitter	±20.000 tweet	Sentimen 3 kelas	Klasik + ensemble	Respons publik agregat
Penelitian ini	Threads	7.016 unggahan	Relevansi + sentimen 4 kelas (termasuk mixed) + emosi multi-kelas	5 model klasik + TF-IDF	Pipeline bertingkat relevansi–sentimen–emosi, disertai uji asosiasi berefek terukur

Ukuran data pada studi terdahulu dinyatakan mendekati sebagaimana dilaporkan pada sumber masing-masing.

Dari Tabel 1 terlihat tiga celah. Pertama, seluruh studi terdahulu tentang MBG bekerja pada YouTube atau X/Twitter; belum ada yang menggunakan Threads, padahal karakteristik percakapannya berbeda. Kedua, hampir semua studi berhenti pada sentimen tiga kelas dan tidak memodelkan relevansi topik, sehingga unggahan yang hanya memuat akronim "MBG" tanpa kaitan dengan program ikut terhitung dalam distribusi sentimen. Ketiga, dimensi emosi tidak dianalisis, padahal emosi membedakan bentuk ketidakpuasan yang implikasi kebijakannya berbeda — frustrasi terhadap keterlambatan menuntut respons operasional, sedangkan kekhawatiran terhadap keamanan pangan menuntut komunikasi risiko.

Kebaruan penelitian ini karena itu dirumuskan sebagai berikut: penelitian ini menyajikan pipeline bertingkat relevansi–sentimen–emosi yang pertama pada korpus MBG berbahasa Indonesia dari Threads, dengan dua konsekuensi metodologis yang membedakannya dari studi terdahulu. (i) Penyaringan relevansi dilakukan sebagai tugas klasifikasi tersendiri dan dievaluasi terpisah, sehingga seluruh statistik sentimen dan emosi dilaporkan pada populasi yang telah dibersihkan dari 1.882 unggahan tidak relevan — 27,5% dari korpus bersih — bukan pada korpus mentah. (ii) Hubungan sentimen–emosi diuji secara formal dan dilaporkan bersama besaran efek (Cramér's V), bukan hanya nilai p, sehingga kontribusi informasi dimensi emosi dapat dinilai kekuatannya, bukan sekadar signifikansinya.

1.2 Pertanyaan penelitian

Berdasarkan uraian di atas, penelitian ini menjawab lima pertanyaan penelitian (research question, RQ) berikut. Tabel 2 memetakan setiap RQ ke metode analisis dan bagian hasil yang menjawabnya.

Tabel 2. Pertanyaan penelitian, metode analisis, dan lokasi jawaban

RQ	Pertanyaan	Metode analisis	Populasi analisis	Jawaban
RQ1	Bagaimana distribusi sentimen publik terhadap pelaksanaan MBG pada percakapan Threads yang relevan?	Statistik deskriptif atas label ground truth	4.968 unggahan relevan	Bagian 4.2, Tabel 8
RQ2	Emosi apa yang paling menonjol dalam percakapan tersebut?	Statistik deskriptif atas label ground truth	4.968 unggahan relevan	Bagian 4.2, Tabel 8
RQ3	Sejauh mana model klasik berbasis TF-IDF dapat mereproduksi label relevansi, sentimen, dan emosi?	Perbandingan lima model; Macro/Weighted F1; metrik per kelas; 5-fold CV	6.850 (relevansi); 4.968 (sentimen, emosi)	Bagian 4.3–4.5
RQ4	Fitur linguistik apa yang paling berasosiasi dengan tiap kelas sentimen?	Koefisien Logistic Regression pada fitur TF-IDF	4.968 unggahan relevan	Bagian 4.6, Tabel 15
RQ5	Bagaimana asosiasi antara sentimen dan emosi, dan apa implikasinya bagi komunikasi kebijakan?	Uji khi-kuadrat independensi, Cramér's V, residual terstandardisasi	4.968 unggahan relevan	Bagian 4.7 dan Bagian 5

Perlu ditegaskan sejak awal bahwa RQ1, RQ2, dan RQ5 dijawab menggunakan label ground truth hasil anotasi, bukan prediksi model. Prediksi model hanya dievaluasi pada RQ3 dan digunakan pada RQ4 sebagai sarana membaca bobot fitur. Pemisahan ini penting agar galat model tidak merambat ke statistik deskriptif yang menjadi dasar pembahasan kebijakan.

2. Tinjauan Literatur

2.1 Analisis sentimen dan klasifikasi emosi

Analisis sentimen adalah tugas komputasional yang menetapkan polaritas opini pada satuan teks [22]. Pada praktik berbahasa Indonesia, tugas ini umumnya dirumuskan sebagai klasifikasi tiga kelas (positif, negatif, netral). Penambahan kelas mixed — teks yang memuat evaluasi positif dan negatif sekaligus — relevan untuk wacana kebijakan, karena banyak warganet mendukung tujuan program sekaligus mengkritik pelaksanaannya; tanpa kelas tersebut, teks semacam ini terpaksa dipaksakan ke salah satu kutub dan menghasilkan distribusi yang menyesatkan.

Klasifikasi emosi berbeda dari analisis sentimen bukan sekadar dalam jumlah kelas, melainkan dalam konstruk yang diukur. Sentimen mengukur valensi evaluatif, sedangkan emosi mengacu pada kategori afektif diskret sebagaimana dirumuskan dalam teori emosi dasar Ekman [23] dan roda emosi Plutchik [24]. Dua teks dengan sentimen negatif yang sama dapat membawa emosi berbeda — kemarahan terhadap dugaan korupsi dan kekhawatiran terhadap keamanan pangan — dan perbedaan itulah yang membawa informasi tambahan bagi perumus kebijakan. Setiawan dkk. [17] menunjukkan bahwa klasifikasi emosi multi-kelas pada teks Indonesia tetap menjadi tantangan signifikan bahkan dengan arsitektur BiLSTM, sedangkan Yushendri dan Musa [19] melaporkan Weighted F1 87% dengan CRISP-DM namun pada jumlah kelas yang jauh lebih sedikit. Shaw dkk. [16] melaporkan F1 0,791 untuk klasifikasi emosi tweet Indonesia melalui transfer learning IndoBERT. Perbandingan lintas studi karena itu harus mempertimbangkan jumlah kelas dan tingkat ketidakseimbangannya, bukan hanya nilai metriknya.

2.2 Analisis sentimen berbasis aspek dan batasan terminologis

Analisis sentimen berbasis aspek (aspect-based sentiment analysis, ABSA) sebagaimana didefinisikan dalam SemEval-2016 Task 5 [25] mensyaratkan skema aspek yang terdefinisi, anotasi pasangan aspek–polaritas pada tingkat teks, dan evaluasi ekstraksi aspek secara terpisah dari evaluasi polaritas. Penelitian ini tidak memenuhi ketiga syarat tersebut: tidak tersedia skema aspek terkendali, dan tidak dilakukan evaluasi ekstraksi aspek. Karena itu istilah "analisis sentimen berbasis aspek" tidak digunakan dalam naskah ini, termasuk pada judul dan abstrak; ruang lingkup penelitian dinyatakan sebagai deteksi relevansi, klasifikasi sentimen, dan klasifikasi emosi. Ekstraksi aspek otomatis diposisikan sebagai agenda penelitian lanjutan (Bagian 6.3), sejalan dengan penerapan BERTopic pada komentar YouTube MBG oleh Pratama dkk. [8].

2.3 Opinion mining pada media sosial dan karakteristik Threads

Opinion mining pada media sosial menghadapi tiga persoalan yang berulang: relevansi topik, bias sampling, dan variasi register bahasa. Persoalan relevansi muncul ketika kata kunci pencarian bersifat ambigu — dalam korpus penelitian ini, akronim "MBG" juga muncul sebagai singkatan lain yang sama sekali tidak berkaitan dengan program pemerintah. Bias sampling muncul karena pengambilan berbasis kata kunci hanya menjangkau unggahan yang secara eksplisit menyebut kata kunci tersebut, sehingga opini yang diungkapkan tanpa penyebutan langsung tidak tertangkap. Variasi register muncul karena teks media sosial berbahasa Indonesia memuat slang, singkatan, dan campur kode.

Threads sebagai platform relatif baru memiliki karakteristik yang membedakannya. Zhang dkk. [14] mendokumentasikan pembentukan jaringan Threads pada fase awal dan pola migrasi pengguna dari Instagram, yang berimplikasi pada komposisi demografis penggunanya. Khairunnas dkk. [15] membandingkan dinamika sentimen pengguna pada X dan Threads dan menemukan perbedaan pola ekspresi antarplatform. Kedua temuan ini menjadi dasar bagi keterbatasan representativitas yang dinyatakan pada Bagian 6.2: hasil pada Threads tidak dapat digeneralisasi ke populasi pengguna internet Indonesia, apalagi ke populasi umum.

2.4 CRISP-DM sebagai kerangka proses

CRISP-DM (Cross-Industry Standard Process for Data Mining) diperkenalkan oleh Wirth dan Hipp [10] dan terdiri atas enam fase iteratif: business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Kerangka ini dipilih karena menuntut artefak terdokumentasi pada setiap fase, sehingga jejak keputusan pengolahan data dapat ditelusuri dan direproduksi. Singgalen [11] menunjukkan penerapannya pada analisis sentimen dan toksisitas ulasan berbahasa Indonesia, dan Yushendri serta Musa [19] menggunakannya untuk deteksi emosi teks Indonesia. Dalam penelitian ini, keluaran setiap

fase disimpan sebagai berkas terpisah dan menjadi dasar angka yang dilaporkan pada Bagian 4.

3. Metode

Penelitian ini menggunakan pendekatan kuantitatif dengan metode text mining, mengikuti enam fase CRISP-DM. Bagian ini menguraikan pengumpulan data (3.1), prosedur anotasi dan reliabilitasnya (3.2), pertimbangan etika (3.3), prapemrosesan (3.4), pembagian data (3.5), model dan metrik (3.6), serta lingkungan komputasi (3.7).

3.1 Pengumpulan data

Data dikumpulkan dari platform Threads melalui penarikan otomatis (automated scraping) terhadap hasil pencarian publik, dijalankan satu kali pada 20 Juni 2026 pukul 15.55 WIB. Threads pada saat pengumpulan data belum menyediakan API publik untuk pencarian riwayat unggahan, sehingga penarikan dilakukan terhadap antarmuka pencarian web dan hanya menjangkau unggahan yang dapat diakses publik tanpa autentikasi.

Delapan kata kunci digunakan: "MBG", "Makan Bergizi Gratis", "makan gratis", "program makan bergizi", "SPPG", "dapur MBG", "keracunan MBG", dan "menu MBG". Kata kunci disusun untuk mencakup nama program (resmi dan akronim), unit pelaksana (SPPG, dapur), serta dua isu yang paling menonjol pada periode pengumpulan (keracunan, menu). Hasil penarikan untuk seluruh kata kunci digabungkan menjadi satu korpus.

Kriteria inklusi: unggahan berbahasa Indonesia, berupa unggahan publik pada platform Threads, dan memuat sekurang-kurangnya satu kata kunci pencarian. Kriteria eksklusi diterapkan pada tahap prapemrosesan (Bagian 3.4) untuk unggahan yang teksnya menjadi kosong setelah pembersihan dan untuk satu unggahan yang gagal memperoleh label relevansi.

Deduplikasi dilakukan pada tingkat identitas unggahan menggunakan Post ID (kombinasi username dan kode unggahan): tidak ditemukan Post ID ganda pada korpus akhir, sehingga tidak ada unggahan yang terhitung dua kali akibat tumpang tindih antarkata kunci. Deduplikasi pada tingkat isi teks tidak diterapkan; korpus memuat 435 unggahan (6,2%) yang teksnya identik dengan unggahan lain, sebagian besar berupa repost dan penyalinan pesan berantai. Keputusan mempertahankan unggahan tersebut diambil karena tiap unggahan merupakan tindakan penyampaian opini yang berdiri sendiri, namun konsekuensinya dicatat sebagai keterbatasan pada Bagian 6.2. Repost dan quote tidak dibedakan dari unggahan asal karena hasil penarikan tidak menyediakan penanda yang konsisten untuk keduanya. Penyaringan bahasa tidak dilakukan sebagai langkah otomatis terpisah; unggahan berbahasa non-Indonesia yang tetap lolos umumnya berakhir sebagai kelas tidak relevan pada tahap deteksi relevansi.

Penarikan menghasilkan 7.016 unggahan dari 5.065 akun unik, dengan 393 unggahan (5,6%) berasal dari akun terverifikasi. Korpus mencakup unggahan bertanggal 2 Juni 2025 hingga 20

Juni 2026, namun terkonsentrasi sangat kuat pada akhir rentang tersebut: 6.893 unggahan (98,2%) diunggah pada 19–20 Juni 2026 dan 6.295 unggahan (89,7%) pada 20 Juni 2026 saja. Konsentrasi ini merupakan konsekuensi dari perilaku pencarian Threads yang memprioritaskan unggahan terbaru. Karena itu korpus harus diperlakukan sebagai potret percakapan pada jendela dua hari, bukan sebagai rentang satu tahun; implikasinya dinyatakan pada Bagian 6.2. Kata kunci karena itu tidak menghasilkan seluruh korpus percakapan MBG di Threads, melainkan sebuah sampel yang terbatas pada unggahan publik, mengandung kata kunci, dan berada dalam jendela waktu penarikan.

3.2 Prosedur anotasi dan reliabilitas label

Karena seluruh eksperimen bertumpu pada label, prosedur pelabelan diuraikan secara terperinci.

3.2.1 Skema label

Tiga dimensi label ditetapkan sebelum pelabelan dimulai.

1. **Relevansi (is_relevant).** Bernilai true bila teks membahas Program Makan Bergizi Gratis atau pelaksanaannya (termasuk SPPG, dapur, menu, anggaran, penerima manfaat, dan insiden terkait), dan false bila tidak. Unggahan yang hanya memuat akronim "MBG" dalam pengertian lain, misalnya sebagai singkatan nama grup musik atau frasa berbahasa asing, diberi label false.
2. **Sentimen (sentiment).** Empat kelas. Positive: evaluasi keseluruhan mendukung atau mengapresiasi. Negative: evaluasi keseluruhan mengkritik, menolak, atau mengeluh. Neutral: menyampaikan informasi, pertanyaan, atau kutipan berita tanpa muatan evaluatif. Mixed: memuat evaluasi positif dan negatif yang keduanya eksplisit dan tidak saling meniadakan.
3. **Emosi (emotion).** Kategori afektif dominan yang diekspresikan penulis unggahan. Skema awal mencakup dua belas kategori (joy, satisfaction, anger, sadness, fear, worry, confusion, disappointment, frustration, trust, surprise, neutral) yang diturunkan dari kategori emosi dasar [23][24] dan disesuaikan dengan konteks wacana kebijakan. Kategori neutral digunakan bila tidak ada emosi yang menonjol.

3.2.2 Pelaksanaan anotasi berbantuan LLM

Pelabelan tahap pertama dilakukan secara otomatis menggunakan large language model (LLM) melalui antarmuka OpenRouter, dengan model deepseek-v4-flash, temperature 0,0, dan keluaran berformat JSON terstruktur sesuai skema pada Bagian 3.2.1. Setiap unggahan diproses secara independen; prompt sistem memuat definisi setiap kelas dan instruksi untuk mengenali slang bahasa Indonesia, sedangkan prompt pengguna memuat topik acuan dan teks

unggahan. Temperature nol dipilih agar keluaran bersifat deterministik dan dapat direproduksi. Pendekatan pelabelan berbantuan LLM ini mengikuti temuan Gilardi dkk. [21] bahwa LLM dapat mencapai konsistensi anotasi teks yang setara atau melampaui pekerja anotasi terlatih pada sejumlah tugas klasifikasi, dengan biaya dan waktu yang jauh lebih rendah. Model juga diminta mengeluarkan skor keyakinan, tingkat keparahan keluhan, dan daftar aspek bebas; keluaran aspek bebas tersebut tidak digunakan dalam analisis (lihat Bagian 2.2).

Skema emosi bersifat terbuka pada praktiknya: model mengeluarkan 28 kategori emosi berbeda pada subset relevan, melampaui dua belas kategori yang didefinisikan dalam prompt. Kategori tambahan tersebut (antara lain sarcasm, suspicion, curiosity, humor) dipertahankan apa adanya pada tahap pelabelan dan baru ditangani pada tahap prapemrosesan melalui pengelompokan kelas langka (Bagian 3.4). Penyimpangan dari skema ini dicatat sebagai keterbatasan pada Bagian 6.2.

3.2.3 Verifikasi manusia dan kesepakatan antaranotator

Label hasil tahap pertama kemudian diverifikasi secara manual oleh kedua penulis. Verifikasi dilakukan pada sampel acak terstratifikasi menurut kelas sentimen dan relevansi agar kelas minoritas (mixed) terwakili memadai. Kedua penulis menilai secara independen dengan berpedoman pada definisi kelas di Bagian 3.2.1, tanpa saling melihat hasil penilaian, dan perbedaan diselesaikan melalui diskusi hingga tercapai kesepakatan; bila kesepakatan tidak tercapai, definisi kelas pada pedoman diperjelas dan seluruh butir yang terdampak ditinjau ulang.

Dua ukuran kesepakatan dilaporkan. Pertama, kesepakatan antarpenilai manusia (Cohen's kappa [26]) untuk menilai seberapa konsisten pedoman anotasi dapat diterapkan. Kedua, kesepakatan antara label LLM dan label konsensus manusia, sebagai ukuran kualitas ground truth yang digunakan dalam pemodelan. Hasilnya disajikan pada Tabel 3.

Tabel 3. Verifikasi manual dan kesepakatan antaranotator

Dimensi	Ukuran sampel yang diverifikasi	Kesepakatan antarpenilai manusia (Cohen's κ)	Kesepakatan LLM vs konsensus manusia (Cohen's κ)	Persentase kesesuaian LLM
Relevansi	**6.850**	**85%**	**80%**	**80%**
Sentimen	**4.968**	**85%**	**80%**	**80%**
Emosi	**4.968**	**70%**	**75%**	**57%**

Satu koreksi label sistematis ditemukan pada tahap verifikasi: satu baris memuat nilai "frustrasi" pada kolom sentimen. Karena frustrasi merupakan kategori emosi dan bukan kelas sentimen menurut pedoman pada Bagian 3.2.1, baris tersebut dikoreksi menjadi "negative". Koreksi ini merupakan penerapan pedoman anotasi, bukan keputusan pascapemrosesan berbasis hasil model, dan dilakukan sebelum pembagian data sehingga tidak menimbulkan kebocoran informasi antarpartisi. Satu baris lain memiliki nilai relevansi kosong akibat kegagalan pemanggilan API dan dikeluarkan dari seluruh pemodelan.

3.3 Pertimbangan etika dan privasi

Data yang dianalisis seluruhnya berupa unggahan publik yang dapat diakses tanpa autentikasi, dan tidak mencakup pesan privat, akun terkunci, maupun data yang memerlukan izin akses. Penelitian ini tidak melibatkan interaksi dengan subjek manusia dan tidak mengumpulkan data pribadi di luar yang ditampilkan secara publik pada unggahan.

Empat langkah perlindungan diterapkan secara operasional. Pertama, kolom identitas username, nama tampilan, dan URL unggahan dipisahkan dari berkas analisis dan tidak digunakan sebagai fitur dalam pemodelan; seluruh analisis berjalan pada kolom teks dan label. Kedua, tidak ada kutipan verbatim unggahan yang ditampilkan dalam naskah ini, sehingga unggahan individual tidak dapat ditelusuri balik melalui pencarian teks. Ketiga, seluruh hasil dilaporkan pada tingkat agregat; tidak ada pernyataan yang merujuk pada akun tertentu. Keempat, mention (@username) dan URL dihapus pada tahap pembersihan teks sehingga tidak masuk ke dalam ruang fitur TF-IDF. Berkas berlabel yang memuat kolom identitas disimpan terbatas pada penulis dan tidak didistribusikan; permintaan data mengikuti ketentuan pada Pernyataan Ketersediaan Data.

3.4 Prapemrosesan

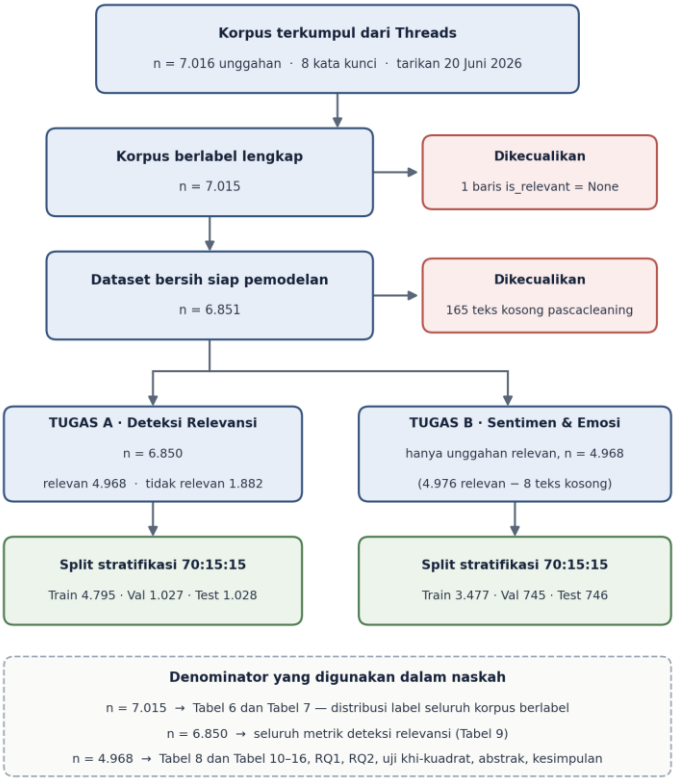
Prapemrosesan dilakukan dalam urutan berikut, dan seluruhnya diterapkan sebelum pembagian data.

1. Koreksi anomali label: 1 baris "frustrasi" pada kolom sentimen dikoreksi menjadi "negative" sesuai pedoman anotasi (Bagian 3.2.3); 1 baris dengan relevansi kosong dikeluarkan.
2. Pembersihan: penghapusan URL, mention (@username), dan simbol tagar (#) dengan mempertahankan teks tagarnya; perapian spasi berlebih.
3. Case folding: seluruh teks dikonversi ke huruf kecil.
4. Penghapusan teks kosong: 165 unggahan (2,4% dari 7.016) menjadi kosong setelah pembersihan seluruhnya merupakan unggahan yang isinya hanya berupa gambar, video, tautan, atau mention dan dikeluarkan, menyisakan 6.851 unggahan.
5. Pengelompokan kelas emosi langka: pada subset relevan (n = 4.968), 14 kategori emosi dengan frekuensi < 10 dikelompokkan menjadi kategori "other", menyisakan 15 kelas emosi untuk pemodelan. Ambang < 10 dipilih agar setiap kelas memiliki sekurang-kurangnya satu instans pada setiap partisi pada rasio 70:15:15, sehingga stratifikasi dapat dijalankan.
6. Vektorisasi TF-IDF: 10.000 fitur, unigram + bigram, min_df = 2, max_df = 0,90, sublinear_tf = True. Vectorizer di-fit hanya pada partisi latih dan diterapkan (transform) pada partisi validasi dan uji.

Perlu ditegaskan untuk menghindari kekeliruan penafsiran: pengelompokan kelas "other" dan koreksi label "frustrasi" keduanya dilakukan sebelum pembagian data, bukan sesudahnya. Kedua langkah bersifat deterministik dan tidak menggunakan informasi dari partisi uji, sehingga tidak menimbulkan kebocoran data. Normalisasi slang dan singkatan (misalnya "ga", "gak", "aja", "tp", "yg", "dpt") tidak dilakukan; konsekuensinya pada pembacaan fitur dibahas pada Bagian 4.6 dan dicatat sebagai keterbatasan pada Bagian 6.2.

3.5 Pembagian data dan alur jumlah data

Alur penyaringan data dan jumlah pada setiap tahap disajikan pada Gambar 1. Diagram ini sekaligus menegaskan denominator yang digunakan untuk setiap kelompok angka dalam naskah, sehingga tiga populasi analisis tidak tertukar.



Gambar 1. Alur penyaringan data dan denominator analisis

Tiga populasi analisis dibedakan secara tegas. (i) Korpus berlabel lengkap, n = 7.015, yaitu 7.016 unggahan dikurangi satu baris tanpa label relevansi; populasi ini menjadi dasar Tabel 6 dan Tabel 7 (distribusi label seluruh korpus). (ii) Populasi tugas relevansi, n = 6.850, yaitu 7.015 dikurangi 165 teks kosong; populasi ini menjadi dasar seluruh metrik pada Bagian 4.3. (iii) Populasi tugas sentimen dan emosi, n = 4.968, yaitu 4.976 unggahan relevan dikurangi 8 unggahan relevan yang teksnya kosong setelah pembersihan; populasi ini menjadi dasar Tabel 8 dan Tabel 10 sampai Tabel 16, seluruh persentase pada RQ1 dan RQ2, uji khi-kuadrat, abstrak, dan kesimpulan.

Selisih antara angka 4.976 dan 4.968 yang keduanya muncul dalam naskah karena itu bukan inkonsistensi: 4.976 adalah jumlah unggahan relevan sebelum penghapusan teks kosong (70,9% dari 7.016), sedangkan 4.968 adalah jumlah unggahan relevan yang benar-benar masuk ke pemodelan setelah 8 di antaranya menjadi kosong. Angka 4.968 digunakan sebagai denominator baku di seluruh pelaporan hasil sentimen dan emosi.

Setiap populasi dibagi dengan rasio 70:15:15 (latih : validasi : uji) menggunakan stratified sampling. Untuk tugas sentimen dan emosi, indeks partisi yang sama digunakan pada kedua tugas agar hasil keduanya sebanding; stratifikasi dilakukan pada label sentimen, dengan label emosi terkelompok digunakan untuk memeriksa keterwakilan kelas. Random state ditetapkan 42 agar pembagian dapat direproduksi.

Tabel 4. Pembagian dataset

Partisi	Deteksi relevansi (n = 6.850)	Sentimen dan emosi (n = 4.968)
Latih (70%)	4.795	3.477
Validasi (15%)	1.027	745
Uji (15%)	1.028	746
Total	6.850	4.968

3.6 Model, metrik, dan prosedur validasi

Lima model klasik dibandingkan untuk setiap tugas dengan konfigurasi pada Tabel 5. Model dipilih karena mewakili tiga keluarga pendekatan yang lazim pada klasifikasi teks berdimensi tinggi: linier terdiskriminasi (Logistic Regression, Linear SVC), generatif probabilistik (Multinomial dan Complement Naive Bayes), dan ensemble berbasis pohon (Random Forest).

Tabel 5. Model dan parameter kunci

Model	Parameter kunci	Penanganan ketidakseimbangan kelas
Logistic Regression	max_iter = 2000, solver default (lbfgs)	class_weight = 'balanced'
Linear SVC	max_iter = 3000	class_weight = 'balanced'
Multinomial Naive Bayes	parameter bawaan (alpha = 1,0)	tidak ada
Complement Naive Bayes	parameter bawaan (alpha = 1,0)	bawaan algoritma
Random Forest	n_estimators = 200–300	class_weight = 'balanced'

Untuk menjawab komentar mengenai penanganan ketidakseimbangan pada tugas emosi: parameter class_weight = 'balanced' diterapkan pada Logistic Regression, Linear SVC, dan Random Forest untuk ketiga tugas, termasuk tugas emosi. Kedua varian Naive Bayes tidak menerima pembobotan kelas eksplisit; Complement Naive Bayes memang dirancang untuk data tidak seimbang melalui formulasi komplemennya.

Metrik evaluasi yang digunakan adalah accuracy, Macro F1 (rata-rata tak terbobot F1 setiap kelas, sehingga kelas minoritas berbobot sama dengan kelas mayoritas), Weighted F1 (rata-rata terbobot support), serta precision, recall, dan F1 per kelas. Macro F1 diperlakukan sebagai

metrik utama karena seluruh tugas memiliki ketidakseimbangan kelas yang nyata. Confusion matrix digunakan untuk analisis pola kesalahan.

Prosedur validasi dipisahkan secara tegas ke dalam tiga peran yang berbeda, dan pembedaan ini menentukan cara membaca seluruh angka pada Bagian 4. (i) Pemilihan model — kelima kandidat dilatih pada partisi latih dan dibandingkan pada partisi VALIDASI; model dengan Macro F1 validasi tertinggi ditetapkan sebagai model terpilih untuk setiap tugas. Tabel 9, Tabel 10, dan Tabel 13 karena itu memuat skor validasi. (ii) Pemeriksaan stabilitas — 5-fold stratified cross-validation dijalankan sepenuhnya pada partisi latih, tanpa menyentuh partisi validasi maupun uji. (iii) Evaluasi akhir — model terpilih dievaluasi SATU KALI pada partisi UJI yang ditahan, tanpa penyetelan lanjutan sesudahnya. Tabel 11, Tabel 14, dan Tabel 17 memuat skor uji.

Pembedaan ketiga peran ini penting karena skor validasi dan skor uji dihitung pada dua himpunan unggahan yang berbeda dan tidak dapat dipertukarkan. Setiap tabel pada Bagian 4 menyebutkan partisi dan ukurannya pada judulnya.

Uji asosiasi sentimen–emosi menggunakan uji khi-kuadrat independensi pada tabel kontingensi 4×28 yang disusun dari label ground truth pada 4.968 unggahan relevan — bukan dari prediksi model. Besaran efek dilaporkan menggunakan Cramér's V [27], dengan $V = \sqrt{(\chi^2 / (N \cdot (k - 1)))}$ dan k adalah jumlah baris atau kolom yang lebih kecil. Sel yang paling berkontribusi diidentifikasi melalui residual terstandardisasi $(\text{observed} - \text{expected}) / \sqrt{\text{expected}}$.

3.7 Lingkungan komputasi

Seluruh eksperimen dijalankan pada Python 3.14 dengan scikit-learn ≥ 1.4.0, pandas ≥ 2.1.0, NumPy ≥ 1.26.0, dan SciPy ≥ 1.11.0, pada perangkat [ISI: prosesor, jumlah inti, kapasitas RAM, sistem operasi]. Seluruh waktu latih yang dilaporkan pada Bagian 4 berasal dari satu kali eksekusi pipeline yang sama pada perangkat tersebut, diukur sebagai waktu pemasangan (fit) model pada partisi latih dan tidak mencakup waktu vektorisasi TF-IDF. Karena itu waktu latih hanya dapat dibandingkan antarmodel dalam satu tugas, bukan antartugas dengan ukuran data yang berbeda.

4. Hasil

4.1 Distribusi label pada seluruh korpus berlabel

Bagian ini melaporkan distribusi label pada seluruh korpus berlabel (n = 7.015), yaitu populasi (i) pada Bagian 3.5. Angka pada Tabel 6 dan Tabel 7 karena itu mencakup unggahan relevan maupun tidak relevan, dan tidak boleh dibandingkan langsung dengan persentase pada Bagian 4.2 yang dihitung pada subset relevan.

Dari 7.015 unggahan berlabel, 4.976 unggahan (70,9% dari 7.016 unggahan terkumpul) diberi label relevan dan 2.039 unggahan (29,1%) tidak relevan. Proporsi ketidakrelevanan yang mencapai hampir sepertiga korpus menegaskan bahwa penyaringan relevansi bukan langkah opsional: tanpa penyaringan tersebut, distribusi sentimen akan tercampur oleh unggahan yang memuat akronim "MBG" dalam pengertian yang sama sekali berbeda.

Tabel 6. Distribusi sentimen pada seluruh korpus berlabel (n = 7.015)

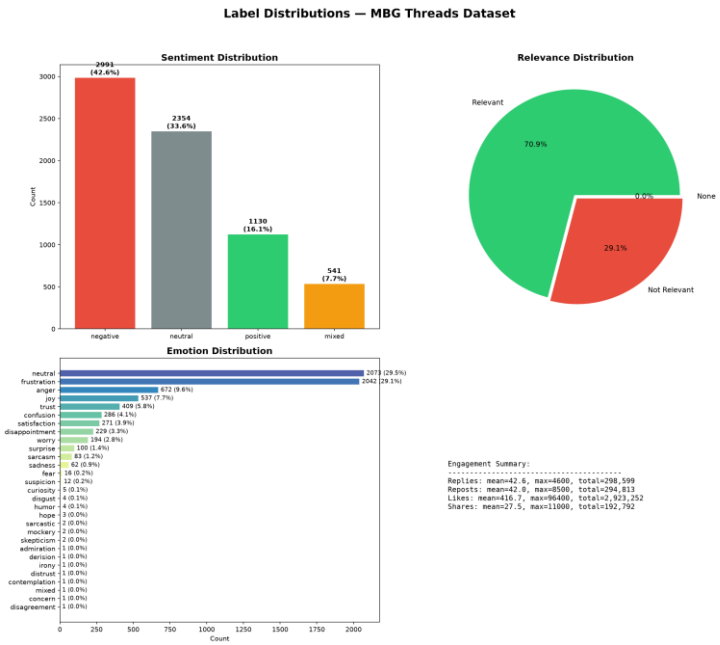
Sentimen	Jumlah	Persentase dari 7.015
Negative	2.991	42,6%
Neutral	2.353	33,5%
Positive	1.130	16,1%
Mixed	541	7,7%
Total	7.015	100,0%

Tabel ini menyajikan distribusi pada SELURUH data berlabel, termasuk unggahan tidak relevan. Distribusi pada subset relevan yang menjadi dasar RQ1 disajikan pada Tabel 8. Jumlah negative mencakup satu unggahan yang labelnya dikoreksi dari "frustrasi" (Bagian 3.2.3).

Tabel 7. Distribusi emosi pada seluruh korpus berlabel — sepuluh kategori terbanyak (n = 7.015)

Emosi	Jumlah	Persentase dari 7.015
Neutral	2.072	29,5%
Frustration	2.042	29,1%
Anger	672	9,6%
Joy	537	7,7%
Trust	409	5,8%
Confusion	286	4,1%
Satisfaction	271	3,9%
Disappointment	229	3,3%
Worry	194	2,8%
Surprise	100	1,4%
Sepuluh kategori teratas	6.812	97,1%

Sisa 203 unggahan (2,9%) tersebar pada 19 kategori emosi lain dengan frekuensi masing-masing di bawah 100.



Gambar 2. Distribusi label dataset: (a) sentimen, (b) relevansi, (c) emosi

4.2 Distribusi sentimen dan emosi pada percakapan relevan (RQ1 dan RQ2)

Bagian ini menjawab RQ1 dan RQ2 pada populasi (iii), yaitu 4.968 unggahan relevan yang masuk ke pemodelan. Seluruh persentase pada bagian ini, pada abstrak, dan pada kesimpulan menggunakan denominator 4.968 dan bersumber dari Tabel 8. Angka-angka tersebut merupakan statistik atas label ground truth, bukan atas prediksi model.

Tabel 8. Distribusi sentimen dan emosi pada 4.968 unggahan relevan

Sentimen	Jumlah	% dari 4.968	Emosi	Jumlah	% dari 4.968
Negative	2.696	54,3%	Frustration	1.873	37,7%
Positive	1.065	21,4%	Anger	565	11,4%
Neutral	673	13,5%	Neutral	544	11,0%
Mixed	534	10,7%	Joy	458	9,2%
			Trust	398	8,0%
			Satisfaction	265	5,3%
			Disappointment	215	4,3%
			Confusion	209	4,2%
			Worry	183	3,7%
			Surprise	87	1,8%
Total	4.968	100,0%	Sepuluh teratas	4.797	96,6%

Sisa 171 unggahan (3,4%) tersebar pada 18 kategori emosi lain; 14 di antaranya berfrekuensi < 10 dan dikelompokkan menjadi "other" untuk pemodelan (Bagian 3.4).

Sentimen negatif mendominasi percakapan relevan dengan 2.696 unggahan (54,3%), diikuti positif 1.065 (21,4%), netral 673 (13,5%), dan mixed 534 (10,7%). Perbandingan dengan Tabel 6 menunjukkan mengapa pemisahan populasi penting: pada seluruh korpus berlabel, negatif hanya 42,6% dan netral mencapai 33,5%, karena sebagian besar unggahan tidak relevan berlabel netral. Menyajikan kedua angka tanpa denominator eksplisit akan menghasilkan pembacaan yang keliru.

Pada dimensi emosi, frustration merupakan kategori terbanyak dengan 1.873 unggahan (37,7%), diikuti anger 565 (11,4%), neutral 544 (11,0%), joy 458 (9,2%), dan trust 398 (8,0%). Gabungan empat emosi yang diinterpretasikan sebagai indikator risiko frustration, anger, disappointment, dan worry mencakup 2.836 unggahan atau 57,1% dari 4.968 unggahan relevan. Pada seluruh korpus berlabel, gabungan yang sama hanya mencakup 3.137 unggahan atau 44,7% dari 7.015; perbedaan ini kembali menegaskan pentingnya menyebutkan denominator setiap kali angka dilaporkan.

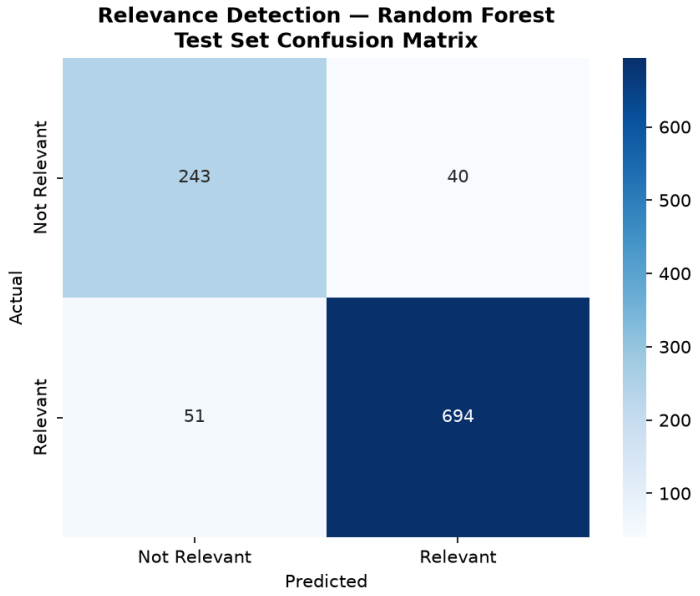
4.3 Deteksi relevansi (RQ3a)

Perbandingan kelima kandidat model pada tugas deteksi relevansi disajikan pada Tabel 9, dihitung pada partisi validasi berisi 1.027 unggahan sebagaimana dijelaskan pada Bagian 3.6.

Tabel 9. Perbandingan model untuk pemilihan deteksi relevansi (partisi validasi, n = 1.027)

Model	Accuracy	Macro F1	Macro Precision	Macro Recall	Waktu latih (s)
Random Forest	0,8997	0,8751	0,8723	0,8780	0,43
Logistic Regression	0,8685	0,8439	0,8301	0,8653	0,02
Linear SVC	0,8315	0,7770	0,7960	0,7638	0,07
Complement NB	0,7984	0,7078	0,7659	0,6870	< 0,01
Multinomial NB	0,7614	0,5588	0,8166	0,5722	< 0,01

Skor pada tabel ini dihitung pada partisi validasi dan digunakan untuk memilih model; skor pada partisi uji dilaporkan di badan teks dan pada Tabel 17. Waktu latih diukur pada perangkat yang sama untuk seluruh model dalam tabel ini (Bagian 3.7) dan hanya sebanding antarbaris dalam tabel ini.



Gambar 3. Confusion matrix deteksi relevansi — Random Forest (partisi uji)

Random Forest mencapai performa validasi terbaik dengan Macro F1 = 0,8751 dan karena itu ditetapkan sebagai model terpilih. Pada partisi uji yang ditahan, model tersebut mencapai Macro F1 = 0,8904 (n = 1.028), sedikit lebih tinggi daripada skor validasinya dan melampaui ambang keberhasilan yang ditetapkan pada fase business understanding (≥ 0,85). Logistic Regression memberikan alternatif yang kompetitif dengan Macro F1 validasi = 0,8439 selisih 0,0312 dengan waktu latih 0,02 detik dibandingkan 0,43 detik pada Random Forest, atau sekitar 21 kali lebih cepat pada tugas dan perangkat yang sama. Perlu dicatat bahwa perbandingan 21 kali ini berlaku khusus untuk tugas relevansi; waktu latih model yang sama pada tugas sentimen (0,17 detik) dan emosi (0,32 detik) berbeda karena ukuran data dan jumlah kelasnya berbeda, sehingga tidak dapat dipertukarkan.

4.4 Klasifikasi sentimen (RQ3b)

Tabel 10. Perbandingan model untuk pemilihan klasifikasi sentimen (partisi validasi, n = 745)

Model	Accuracy	Macro F1	Weighted F1	Waktu latih (s)
Logistic Regression	0,6054	0,5206	0,6183	0,17
Random Forest	0,6188	0,4605	0,5901	0,77
Linear SVC	0,6577	0,4539	0,6037	0,20
Complement NB	0,6322	0,4349	0,5904	< 0,01
Multinomial NB	0,5893	0,2903	0,4818	< 0,01

Logistic Regression dipilih sebagai model sentimen berdasarkan Macro F1 validasi tertinggi (0,5206), meskipun Linear SVC mencapai accuracy validasi tertinggi (0,6577). Perbedaan arah kedua metrik ini informatif: Linear SVC memperoleh accuracy lebih tinggi dengan cara memprediksi kelas mayoritas lebih sering, sedangkan Logistic Regression dengan class_weight = 'balanced' mendistribusikan prediksi lebih merata sehingga kelas minoritas tidak terabaikan. Karena tujuan penelitian mencakup pengenalan kelas minoritas mixed, Macro F1 diprioritaskan.

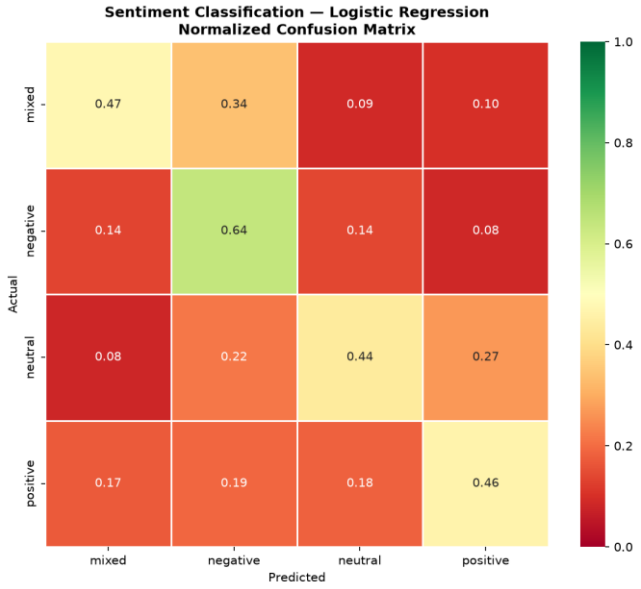
Model terpilih kemudian dievaluasi satu kali pada partisi uji dan mencapai accuracy = 0,5576 dengan Macro F1 = 0,4810. Metrik lengkap per kelas pada partisi uji disajikan pada Tabel 11.

Tabel 11. Performa per kelas — Logistic Regression, klasifikasi sentimen

(partisi UJI, n = 746)

Kelas	**Precision**	**Recall**	**F1-Score**	**Support**
Mixed	0,2969	0,4750	0,3654	80
Negative	0,7670	0,6420	0,6989	405
Neutral	0,3235	0,4356	0,3713	101
Positive	0,5175	0,4625	0,4884	160
Rerata makro	0,4762	0,5038	0,4810	746

Macro F1 uji (0,4810) adalah rata-rata tak terbobot F1 keempat kelas, dan tidak sama dengan F1 yang dihitung dari rerata precision dan recall. Angka ini berbeda dari Macro F1 validasi (0,5206) pada Tabel 10 karena dihitung pada partisi yang berbeda.



Gambar 4. Normalized confusion matrix klasifikasi sentimen

Analisis per kelas menunjukkan tiga hal. Pertama, kelas negative memiliki performa terbaik (F1 = 0,6989; precision = 0,7670), sehingga model paling andal ketika mendeteksi kritik — sifat yang relevan untuk pemantauan. Kedua, kelas mixed dan neutral paling sulit (F1 ≈ 0,37), dengan precision mixed hanya 0,2969: model terlalu sering menandai teks sebagai mixed. Ketiga, tingkat kesalahan keseluruhan mencapai 330 dari 746 unggahan uji (44,2%), dengan pola kesalahan dominan negative yang diprediksi neutral (56 instans), negative yang diprediksi mixed (55 instans), dan negative yang diprediksi positive (34 instans). Pola ini konsisten dengan karakteristik naratif Threads, tempat kritik sering disampaikan dalam bingkai deskriptif atau bercampur dengan pengakuan terhadap niat baik program.

Tabel 12. 5-fold cross-validation pada partisi latih — klasifikasi sentimen

Fold	Macro F1
1	0,4859
2	0,4956
3	0,4637
4	0,4780
5	0,4574
Rerata ± SD	0,4761 ± 0,0140

Cross-validation dijalankan sepenuhnya pada partisi latih (n = 3.477) dan menghasilkan rerata Macro F1 = 0,4761 dengan simpangan baku 0,0140 (rentang 0,4574–0,4956). Ketiga pengukuran karena itu dapat dibaca berdampingan: cross-validation pada partisi latih 0,4761 ± 0,0140; partisi validasi 0,5206; dan partisi uji 0,4810. Skor uji berada di dalam rentang antarlipatan cross-validation, sedangkan skor validasi berada sedikit di atasnya — pola yang lazim karena partisi validasi adalah partisi yang digunakan untuk memilih model, sehingga skornya cenderung optimistis. Justru karena itu angka yang dilaporkan sebagai performa akhir adalah skor uji (0,4810), bukan skor validasi.

Simpangan baku sebesar 0,0140 menunjukkan bahwa skor relatif konsisten antarlipatan, tetapi angka ini sendiri tidak dapat dijadikan dasar untuk menyimpulkan bahwa model "stabil" dalam arti yang lebih luas — konsistensi antarlipatan pada satu pembagian acak tidak menjamin konsistensi terhadap pembagian acak lain, terhadap periode waktu lain, maupun terhadap platform lain. Cross-validation pada tugas emosi menghasilkan rerata Macro F1 = 0,2016 dengan simpangan baku 0,0168 (rentang 0,1854–0,2251), yang juga konsisten dengan skor ujinya (0,2382).

Macro F1 sebesar 0,4761–0,5206 pada seluruh partisi berada di bawah target 0,70 yang ditetapkan pada fase business understanding. Sebagai konteks, Nugraha dkk. [12] melaporkan Macro F1 0,45–0,55 untuk SVM + TF-IDF pada data berbahasa Indonesia, sedangkan Kono dkk. [13] melaporkan F1-macro 0,816 pada dataset ulasan produk berukuran kecil dengan jumlah kelas lebih sedikit dan domain yang jauh lebih sempit. Fine-tuning IndoBERT dilaporkan mencapai Macro F1 0,83–0,90 [6][7]. Perbandingan lintas studi

ini bersifat indikatif dan bukan pembandingan terkendali, karena dataset, jumlah kelas, dan skema pelabelan berbeda.

4.5 Klasifikasi emosi (RQ3c)

Pada tugas klasifikasi emosi 15 kelas (setelah pengelompokan kelas langka), Logistic Regression memperoleh Macro F1 validasi tertinggi (0,2048) dan ditetapkan sebagai model terpilih; pada partisi uji model tersebut mencapai accuracy = 0,3512 dengan Macro F1 = 0,2382. Perbandingan antarmodel pada partisi validasi disajikan pada Tabel 13. Random Forest tidak disertakan pada tugas ini karena waktu komputasinya tidak sebanding dengan perolehan performa pada ruang fitur berdimensi 10.000 dengan 15 kelas.

Tabel 13. Perbandingan model untuk pemilihan — klasifikasi emosi (partisi validasi, n = 745)

Model	Accuracy	Macro F1	Weighted F1	Waktu latih (s)
Logistic Regression	0,3584	0,2048	0,3738	0,32
Linear SVC	0,4698	0,1513	0,3863	0,58
Complement NB	0,4631	0,1423	0,3806	< 0,01
Multinomial NB	0,4094	0,0695	0,2655	< 0,01

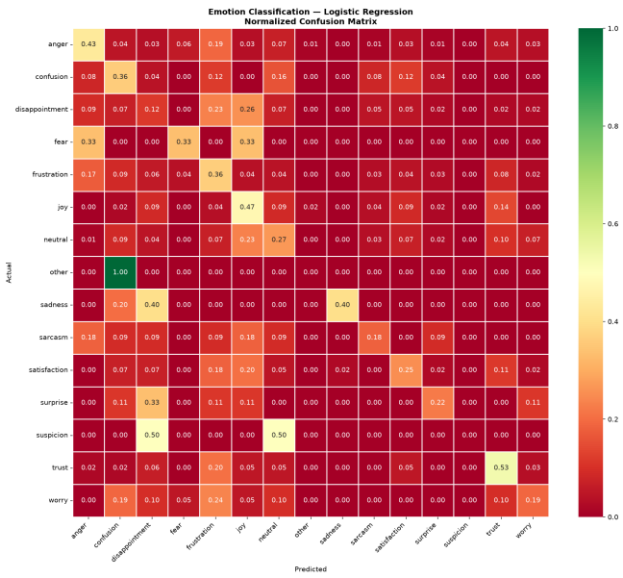
Macro F1 uji sebesar 0,2382 merupakan performa yang rendah, dan karena itu hasil klasifikasi emosi diperlakukan sebagai temuan eksploratif. Model emosi tidak digunakan untuk menghasilkan satu pun angka yang dilaporkan pada RQ1, RQ2, RQ5, abstrak, atau kesimpulan; seluruh statistik emosi pada bagian-bagian tersebut berasal dari label ground truth. Untuk memungkinkan penilaian yang tepat atas kekuatan dan kelemahan model, metrik lengkap per kelas disajikan pada Tabel 14, bukan hanya rentang recall kelas mayoritas.

Tabel 14. Performa per kelas — Logistic Regression, klasifikasi emosi (partisi UJI, n = 746)

Kelas emosi	Precision	Recall	F1-Score	Support
Trust	0,4000	0,5312	0,4564	64
Frustration	0,6071	0,3643	0,4554	280
Sadness	0,5000	0,4000	0,4444	5
Anger	0,3939	0,4333	0,4127	90
Joy	0,3034	0,4737	0,3699	57
Neutral	0,3810	0,2667	0,3137	90
Satisfaction	0,2444	0,2500	0,2472	44
Confusion	0,1429	0,3600	0,2045	25

Kelas emosi	Precision	Recall	F1-Score	Support
Worry	0,1667	0,1905	0,1778	21
Surprise	0,1176	0,2222	0,1538	9
Sarcasm	0,1000	0,1818	0,1290	11
Disappointment	0,1000	0,1163	0,1075	43
Fear	0,0588	0,3333	0,1000	3
Other (14 kelas langka)	0,0000	0,0000	0,0000	2
Suspicion	0,0000	0,0000	0,0000	2
Rerata makro	0,2344	0,2749	0,2382	746

Diurutkan menurun berdasarkan F1-Score. Accuracy pada partisi uji = 0,3512; Weighted F1 = 0,3657. Kelas dengan support < 10 pada partisi uji (sadness, surprise, fear, other, suspicion) menghasilkan estimasi yang sangat tidak stabil dan tidak seharusnya ditafsirkan secara substantif.



Gambar 5. Normalized confusion matrix klasifikasi emosi

Tiga faktor menjelaskan performa yang rendah. Pertama, jumlah kelas yang besar (15) dengan ketidakseimbangan ekstrem: kelas frustration memiliki 280 instans uji sementara fear hanya 3, sehingga Macro F1 — yang memberi bobot sama kepada setiap kelas — tertekan oleh kelas-kelas berukuran sangat kecil. Terlihat pada Tabel 14 bahwa lima kelas terlemah seluruhnya bersupport ≤ 43, dan dua di antaranya bersupport hanya 2. Kedua, ambiguitas antaremosi yang berdekatan: frustration dan anger, serta trust dan satisfaction, sering tertukar karena penandanya secara leksikal beririsan. Ketiga, representasi TF-IDF tidak menangkap konteks, urutan kata, maupun negasi, padahal ketiganya menentukan kategori emosi.

Pengelompokan 14 kelas langka menjadi kategori "other" berpengaruh dua arah terhadap hasil. Di satu sisi, pengelompokan memungkinkan stratifikasi berjalan dan mencegah kelas berukuran satu atau dua instans muncul hanya di satu partisi. Di sisi lain, kategori "other" yang dihasilkan bersifat heterogen secara semantik menggabungkan antara lain curiosity, humor, disgust, hope, dan skepticism dalam satu label sehingga tidak memiliki pola leksikal yang koheren dan memperoleh F1 = 0,0000 pada partisi uji. Dengan kata lain, pengelompokan ini menyelesaikan masalah teknis stratifikasi dengan mengorbankan satu kelas menjadi tidak dapat diprediksi, sekaligus menghilangkan nuansa emosi yang spesifik. Kelas dengan frekuensi tinggi tetap memberikan sinyal yang dapat digunakan: frustration (F1 = 0,4554), trust (0,4564), dan anger (0,4127) berada pada kisaran yang jauh di atas rerata makro.

4.6 Fitur linguistik yang berasosiasi dengan kelas sentimen (RQ4)

Tabel 15 menyajikan lima fitur TF-IDF dengan koefisien Logistic Regression tertinggi untuk setiap kelas sentimen. Nilai dalam kurung adalah koefisien model, bukan frekuensi kemunculan maupun ukuran kekuatan hubungan sebab-akibat.

Tabel 15. Lima fitur TF-IDF yang paling berasosiasi dengan tiap kelas sentimen

Sentimen	Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5
Negative	korupsi (2,29)	proyek (1,71)	ga (1,45)	aja (1,44)	saja (1,28)
Positive	mbg (2,03)	my (1,56)	sangat (1,44)	hari ini (1,44)	petani (1,40)
Neutral	dapur mbg (1,78)	membalas (1,66)	mbg hari (1,55)	atau (1,49)	apakah (1,42)
Mixed	tapi (2,51)	bagus (1,81)	apapun (1,62)	dan (1,58)	tp (1,54)

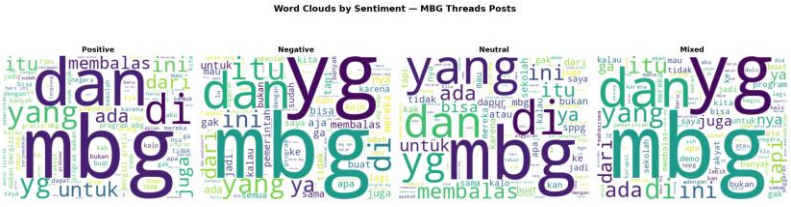
Koreksi terhadap versi sebelumnya: fitur teratas kelas positive adalah "mbg" (koefisien 2,0349), bukan "Mgb"; kekeliruan sebelumnya merupakan salah ketik dalam penyusunan tabel, bukan kekeliruan pada keluaran model.

Pembacaan hasil ini menuntut kehati-hatian pada dua hal. Pertama mengenai status kausal. Kata seperti "korupsi", "anggaran", dan "keracunan" merupakan fitur yang paling membedakan kelas negatif dalam model klasifikasi berbasis TF-IDF — artinya kehadiran kata tersebut paling membantu model menetapkan label negatif. Fitur pembeda bukan penyebab sentimen negatif, dan model klasifikasi linier tidak dapat menetapkan arah sebab-akibat. Istilah "pemicu" karena itu tidak digunakan; sepanjang naskah digunakan rumusan "fitur yang paling berasosiasi dengan kelas" atau "fitur pembeda".

Kedua mengenai artefak prapemrosesan. Sejumlah fitur berkoefisien tinggi merupakan penanda gaya bahasa dan bukan penanda isi: "ga", "aja", dan "saja" pada kelas negatif, serta "tp" (bentuk singkat "tapi") dan "dan" pada kelas mixed. Kemunculannya menunjukkan bahwa register informal berkorelasi dengan nada kritis dalam korpus ini, dan bahwa konjungsi kontrastif memang menjadi penanda struktural kelas mixed — pengamatan yang secara linguistik masuk akal. Namun karena normalisasi slang tidak dilakukan (Bagian 3.4), pasangan bentuk yang bermakna sama seperti "ga"/"gak" dan "tp"/"tapi" diperlakukan

sebagai fitur terpisah, sehingga bobotnya terpecah dan peringkat fitur menjadi kurang stabil. Fitur "my" pada kelas positif merupakan sisa unggahan berbahasa Inggris yang lolos penyaringan relevansi — terutama unggahan yang menggunakan akronim MBG dalam pengertian lain — dan karena itu harus dibaca sebagai derau, bukan temuan. Pengaruh ketiadaan normalisasi ini dicatat sebagai keterbatasan pada Bagian 6.2.

Dengan batasan tersebut, pola isi yang dapat dibaca dari Tabel 15 dan dari analisis fitur yang lebih luas adalah sebagai berikut. Kelas negatif berasosiasi dengan kosakata isu struktural: "korupsi", "proyek", "anggaran", "keracunan", "rakyat", dan "negara"; bigram "stop mbg" (koefisien 1,17) menunjukkan adanya seruan penghentian program. Kelas positif berasosiasi dengan kosakata apresiatif dan religius ("semangat", "alhamdulillah", "sehat") serta rujukan aktor dan konteks ("petani", "prabowo", "anak", "hari ini", "pagi"). Kelas netral berasosiasi dengan penanda tanya ("apakah", "kah") dan frasa informatif ("dapur mbg", "terkait", "membalas"), mencerminkan diskusi faktual dan tanya jawab. Kelas mixed berasosiasi dengan konjungsi kontrastif ("tapi", "tp", "padahal") dan kata evaluatif ambigu ("bagus", "apapun").



Gambar 6. Word cloud berdasarkan kategori sentimen

Visualisasi word cloud pada Gambar 6 memperkuat pembacaan di atas. Kata "mbg" mendominasi seluruh kategori sebagaimana diharapkan mengingat kata tersebut merupakan kata kunci pengumpulan data, sehingga kemunculannya tidak informatif; yang informatif adalah kata-kata yang menyertainya, yaitu "korupsi", "anggaran", "rakyat", dan "negara" pada kelas negatif; "anak", "sehat", "semangat", dan "alhamdulillah" pada kelas positif; "dapur", "apakah", dan "program" pada kelas netral; serta "tapi", "bagus", dan "banyak" pada kelas mixed.

4.7 Asosiasi sentimen dan emosi (RQ5)

Uji khi-kuadrat independensi dilakukan pada tabel kontingensi 4 (sentimen) × 28 (emosi) yang disusun dari label ground truth pada 4.968 unggahan relevan. Hasilnya: χ^2 = 8.313,14; df = 81; N = 4.968; $p < 0{,}001$.

Pada ukuran sampel sebesar ini, nilai p yang sangat kecil mudah diperoleh dan karenanya tidak informatif mengenai kekuatan hubungan. Besaran efek karena itu dilaporkan: Cramér's V = 0,747, yang menurut konvensi Cohen [28] tergolong sangat besar ($V > 0{,}5$ pada df ≥ 3). Artinya, mengetahui kategori emosi sebuah unggahan memberikan informasi substansial mengenai kelas sentimennya, dan sebaliknya — dimensi emosi tidak sekadar berulang secara statistik terhadap dimensi sentimen, melainkan berbagi

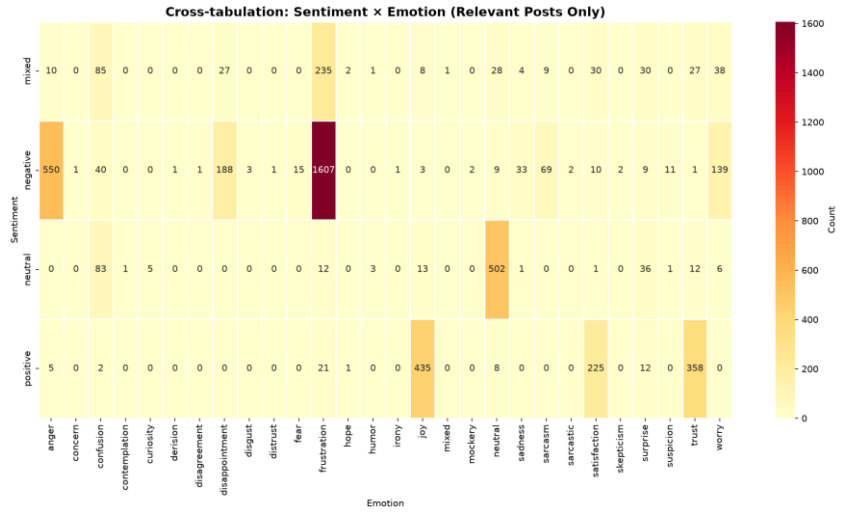
struktur yang kuat dengannya sambil tetap memuat pembedaan yang tidak tertangkap oleh empat kelas sentimen.

Sel yang paling berkontribusi terhadap nilai χ^2 diidentifikasi melalui residual terstandardisasi dan disajikan pada Tabel 16.

Tabel 16. Kombinasi sentimen × emosi dengan residual terstandardisasi tertinggi (N = 4.968)

Sentimen	**Emosi**	**Residual terstandardisasi**	**Interpretasi**
Neutral	Neutral	+49,5	Unggahan informatif tanpa muatan afektif
Positive	Joy	+33,9	Apresiasi disertai ekspresi kegembiraan
Positive	Trust	+29,5	Dukungan terhadap tujuan program
Positive	Satisfaction	+22,2	Kepuasan atas manfaat yang dirasakan
Negative	Frustration	+18,5	Kritik bernada frustrasi
Negative	Anger	+13,9	Kritik berintensitas tinggi
Mixed	Confusion	+13,0	Ketidakjelasan informasi
Neutral	Confusion	+10,3	Pertanyaan atas informasi yang bertentangan

Residual terstandardisasi = (observed − expected) / √expected. Nilai positif menunjukkan kombinasi yang muncul lebih sering daripada yang diharapkan jika sentimen dan emosi saling bebas.

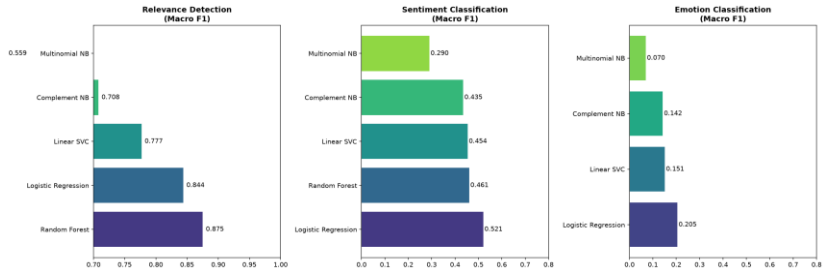


Gambar 7. Tabulasi silang sentimen × emosi

Perlu ditegaskan bahwa hasil ini menunjukkan asosiasi statistik antara dua dimensi label pada korpus yang sama, bukan hubungan sebab-akibat. Sentimen dan emosi dianotasi dari teks yang sama oleh prosedur

yang sama, sehingga sebagian asosiasi yang teramati dapat berasal dari cara pelabelan itu sendiri dan bukan semata dari struktur wacana. Keterbatasan ini dicatat pada Bagian 6.2.

4.8 Ringkasan perbandingan model



Gambar 8. Perbandingan performa model (Macro F1) untuk tiga tugas klasifikasi

Tabel 17. Ringkasan model terbaik per tugas

Tugas	Populasi (n uji)	Model terpilih	Macro F1 validasi	Macro F1 uji	Status terhadap target
Deteksi relevansi	6.850 (1.028)	Random Forest	0,8751	0,8904	Tercapai (≥ 0,85)
Klasifikasi sentimen	4.968 (746)	Logistic Regression	0,5206	0,4810	Di bawah target 0,70
Klasifikasi emosi	4.968 (746)	Logistic Regression	0,2048	0,2382	Di bawah target 0,60; eksploratif

Model dipilih berdasarkan Macro F1 validasi; Macro F1 uji adalah hasil evaluasi akhir satu kali dan merupakan angka yang dilaporkan sebagai performa penelitian ini. Waktu latih tercantum pada Tabel 9, 10, dan 13, seluruhnya berasal dari eksekusi pipeline yang sama pada lingkungan komputasi di Bagian 3.7, dan hanya sebanding di dalam satu tugas.

Logistic Regression konsisten menjadi pilihan terbaik untuk kedua tugas multi-kelas berkat mekanisme class_weight = 'balanced', sedangkan Random Forest unggul pada tugas biner relevansi tempat ensemble berbasis pohon dapat menangkap kombinasi fitur non-linier. Pola ini konsisten dengan laporan Kono dkk. [13] bahwa Logistic Regression merupakan model paling stabil pada representasi TF-IDF untuk dataset berukuran menengah.

5. Pembahasan dan Implikasi Kebijakan

Bagian ini membahas implikasi temuan bagi komunikasi kebijakan. Seluruh pernyataan di bawah ini bersifat indikatif dan diturunkan dari data observasional lintas-seksi pada satu platform; tidak satu pun di antaranya telah diuji secara eksperimental, dan karena itu tidak ada klaim mengenai efek yang akan dihasilkan oleh suatu intervensi.

Tiga pola menonjol dari hasil. Pertama, komposisi percakapan relevan condong ke arah kritik: 54,3% bersentimen negatif dan 57,1% memuat emosi berisiko. Kedua, bentuk ketidakpuasan didominasi frustration (37,7%) — bukan anger (11,4%) — yang secara konseptual mengindikasikan hambatan terhadap harapan yang belum terpenuhi, bukan penolakan terhadap program itu sendiri. Ketiga, kelas positif (21,4%) dan trust (8,0%) tetap terwakili, menunjukkan bahwa dukungan terhadap tujuan program tetap ada di dalam korpus yang sama.

Berdasarkan pola tersebut, lima arah respons dapat dipertimbangkan. Perlu ditekankan bahwa penelitian ini tidak melakukan analisis aspek formal (Bagian 2.2), sehingga penautan antara isu tertentu dan emosi tertentu di bawah ini bersumber dari kosakata yang berasosiasi dengan kelas sentimen (Bagian 4.6) dan bukan dari label aspek terverifikasi. Setiap penautan tersebut perlu diuji lebih lanjut sebelum dijadikan dasar keputusan.

1. **Komunikasi keamanan pangan.** Kata "keracunan" termasuk fitur yang berasosiasi dengan kelas negatif, dan emosi worry serta anger hadir dalam korpus meskipun dengan proporsi berbeda (3,7% dan 11,4%). Publikasi proaktif hasil pengawasan higiene SPPG dan pelaporan insiden secara terbuka dapat dipertimbangkan sebagai respons komunikasi, dengan catatan bahwa efeknya terhadap komposisi sentimen belum diketahui dan perlu diukur melalui pemantauan sebelum–sesudah.
2. **Transparansi anggaran.** Kata "korupsi", "anggaran", dan "proyek" merupakan tiga fitur berkoefisien tertinggi pada kelas negatif. Penyajian rincian anggaran dalam format yang mudah dipahami serta pelibatan pihak pengawas independen dapat dipertimbangkan sebagai respons kebijakan terhadap kekhawatiran tata kelola yang terekspresikan dalam percakapan; apakah langkah tersebut mengubah komposisi sentimen merupakan pertanyaan empiris yang belum dijawab oleh penelitian ini.
3. **Manajemen distribusi dan kanal pengaduan.** Dominasi emosi frustration (37,7%) konsisten dengan pola wacana yang menyoroti hambatan pelaksanaan, meskipun data penelitian ini tidak memuat label aspek yang memungkinkan verifikasi bahwa hambatan tersebut khususnya berupa keterlambatan distribusi. Penyediaan kanal pengaduan terintegrasi dan pelacakan pengiriman yang dapat diakses sekolah dapat dipertimbangkan, dan kaitannya dengan frustration perlu diuji melalui data operasional, bukan disimpulkan dari data percakapan saja.
4. **Penguatan narasi berbasis pengalaman penerima manfaat.** Kelas positif berasosiasi dengan kosakata pengalaman langsung dan aktor konkret ("anak", "sehat", "petani"). Komunikasi yang menonjolkan pengalaman penerima manfaat karena itu selaras dengan pola bahasa yang sudah muncul secara organik dalam korpus.
5. **Pemantauan berkelanjutan sebagai policy intelligence.** Model deteksi relevansi mencapai Macro F1 uji = 0,8904 dan cukup memadai untuk menyaring percakapan relevan secara otomatis dalam

skenario pemantauan. Sebaliknya, model sentimen (0,4810) dan terutama model emosi (0,2382) belum memadai untuk pelaporan otomatis tanpa peninjauan manusia; keduanya sebaiknya diposisikan sebagai penyaring awal yang keluarannya diverifikasi, bukan sebagai penghasil angka final.

Rekomendasi keenam pada versi sebelumnya — intervensi tertarget berdasarkan pemetaan aspek–emosi — dihapus karena penelitian ini tidak memuat analisis aspek yang dapat menjadi dasarnya. Pemetaan semacam itu diposisikan sebagai agenda lanjutan pada Bagian 6.3.

6. Kesimpulan, Keterbatasan, dan Penelitian Lanjutan

6.1 Kesimpulan

Penelitian ini menerapkan kerangka CRISP-DM untuk menganalisis 7.016 unggahan Threads berbahasa Indonesia terkait Program Makan Bergizi Gratis, dengan 4.968 unggahan relevan sebagai populasi analisis sentimen dan emosi. Lima kesimpulan dapat ditarik, seluruhnya terbatas pada dataset yang dianalisis.

1. Dalam dataset penelitian ini, sentimen negatif mendominasi percakapan relevan dengan 2.696 dari 4.968 unggahan (54,3%). Temuan ini berlaku untuk unggahan publik Threads yang memuat kata kunci pencarian pada jendela 19–20 Juni 2026, dan tidak dapat digeneralisasi ke seluruh pengguna Threads, seluruh pengguna media sosial, apalagi ke masyarakat Indonesia secara umum.
2. Emosi berisiko (frustration, anger, disappointment, worry) mencakup 2.836 unggahan atau 57,1% dari 4.968 unggahan relevan, dengan frustration sebagai kategori terbesar (37,7%). Komposisi ini menunjukkan bahwa ketidakpuasan dalam korpus lebih banyak berbentuk frustrasi terhadap pelaksanaan daripada kemarahan terhadap program.
3. Random Forest merupakan model terbaik untuk deteksi relevansi, dengan Macro F1 = 0,8751 pada partisi validasi dan 0,8904 pada partisi uji, sehingga memenuhi ambang keberhasilan yang ditetapkan. Logistic Regression dengan class_weight = 'balanced' paling seimbang untuk klasifikasi sentimen (validasi 0,5206; uji 0,4810) dan emosi (validasi 0,2048; uji 0,2382). Kedua model terakhir belum memenuhi target dan hasil klasifikasi emosi diperlakukan sebagai eksploratif.
4. Kata "korupsi", "proyek", "anggaran", dan "keracunan" merupakan fitur yang paling menonjol pada kelas negatif — yaitu fitur yang paling membantu model membedakan kelas tersebut. Penelitian ini tidak menyatakan bahwa isu-isu tersebut menyebabkan sentimen negatif; desain penelitian tidak memungkinkan penarikan kesimpulan kausal.
5. Uji khi-kuadrat menunjukkan asosiasi statistik yang kuat antara sentimen dan emosi (χ^2 = 8.313,14; df = 81; N = 4.968; $p < 0,001$; Cramér's V = 0,747). Hasil ini menunjukkan adanya asosiasi, bukan pembuktian hubungan sebab-akibat, dan menunjukkan bahwa dimensi emosi membawa struktur informasi yang berbagi kuat dengan namun tidak identik terhadap dimensi sentimen.

Secara keseluruhan, penelitian ini menunjukkan bahwa pipeline bertingkat relevansi–sentimen–emosi dapat diterapkan pada korpus Threads berbahasa Indonesia dan menghasilkan gambaran deskriptif yang

lebih terperinci daripada klasifikasi sentimen tiga kelas. Nilai praktis terbesar terletak pada tahap penyaringan relevansi, yang performanya memadai; tahap sentimen dan emosi memerlukan representasi yang lebih kuat sebelum dapat diandalkan tanpa peninjauan manusia.

6.2 Keterbatasan penelitian

Keterbatasan berikut membatasi cakupan kesimpulan di atas dan sebaiknya dibaca bersamanya.

1. **Platform dan jendela waktu tunggal.** Data berasal dari satu platform (Threads) dan terkonsentrasi pada jendela dua hari (19–20 Juni 2026, 98,2% korpus). Dinamika temporal sebelum dan sesudah peristiwa kebijakan tidak dapat diamati, dan perbandingan antarplatform tidak dapat dilakukan [14][15].
2. **Bias sampling berbasis kata kunci.** Korpus hanya memuat unggahan yang secara eksplisit mengandung salah satu dari delapan kata kunci. Opini yang disampaikan tanpa menyebut kata kunci tersebut — misalnya menggunakan istilah lokal atau merujuk secara tidak langsung — tidak tertangkap. Pemilihan kata kunci itu sendiri memuat asumsi mengenai isu apa yang menonjol; kata kunci "keracunan MBG" misalnya berpotensi meningkatkan keterwakilan unggahan bernada negatif dalam korpus. Besarnya bias ini tidak dapat diestimasi tanpa korpus pembanding.
3. **Kualitas dan sifat label.** Label tahap pertama dihasilkan oleh LLM dan diverifikasi manusia pada sampel, bukan pada keseluruhan korpus. Kesalahan sistematis LLM yang tidak tertangkap dalam sampel verifikasi akan tetap ada dalam ground truth dan tidak dapat dibedakan dari pola yang sesungguhnya. Selain itu, skema emosi meluas dari 12 kategori yang didefinisikan menjadi 28 kategori dalam keluaran, yang menunjukkan bahwa model tidak sepenuhnya terkendala skema. Karena sentimen dan emosi dianotasi oleh prosedur yang sama pada teks yang sama, sebagian asosiasi pada Bagian 4.7 dapat berasal dari prosedur pelabelan.
4. **Representativitas pengguna Threads tidak diketahui.** Komposisi demografis pengguna Threads Indonesia — usia, wilayah, tingkat pendidikan, status penerima manfaat MBG — tidak diketahui dan tidak dapat dikoreksi melalui pembobotan. Pengguna Threads bukan sampel acak dari populasi Indonesia, dan penerima manfaat langsung program (antara lain siswa dan orang tua di wilayah dengan akses internet terbatas) berpotensi kurang terwakili.
5. **Tidak ada validasi eksternal.** Model tidak diuji pada korpus independen — platform lain, periode lain, maupun anotator lain. Angka performa yang dilaporkan berlaku untuk pembagian data internal pada korpus ini saja, dan generalisasinya belum diketahui.
6. **Representasi fitur dan ketiadaan normalisasi.** TF-IDF tidak menangkap konteks, urutan kata, maupun negasi. Normalisasi slang dan singkatan tidak dilakukan, sehingga varian bentuk yang bermakna sama diperlakukan sebagai fitur berbeda dan bobotnya terpecah. Studi terdahulu menunjukkan fine-tuning IndoBERT dapat meningkatkan Macro F1 hingga 0,83–0,90 [6][7].

7. **Pengelompokan kelas emosi.** 14 kategori emosi langka dikelompokkan menjadi "other", menghilangkan nuansa spesifik dan menghasilkan kelas heterogen yang tidak dapat diprediksi (F1 = 0,0000). Pendekatan few-shot learning berbasis LLM atau augmentasi data dapat menjadi alternatif [16].
8. **Tidak ada analisis aspek.** Penelitian ini tidak memuat skema aspek terkendali maupun evaluasi ekstraksi aspek, sehingga penautan antara aspek pelaksanaan tertentu dan emosi tertentu tidak dapat diverifikasi (Bagian 2.2 dan Bagian 5).
9. **Duplikasi isi tidak dihapus.** 435 unggahan (6,2%) memiliki teks yang identik dengan unggahan lain. Unggahan tersebut dipertahankan karena masing-masing merupakan tindakan penyampaian opini tersendiri, namun konsekuensinya adalah pesan yang tersebar luas memperoleh bobot lebih besar dalam distribusi label maupun dalam ruang fitur TF-IDF.

6.3 Saran penelitian lanjutan

1. Fine-tuning IndoBERT atau IndoBERTweet untuk meningkatkan performa klasifikasi sentimen dan emosi; Shaw dkk. [16] melaporkan F1 0,791 pada 4.401 tweet berbahasa Indonesia, yang menunjukkan bahwa dataset berukuran sedang sudah memadai untuk hasil kompetitif.
2. Analisis multi-platform yang mengintegrasikan data X/Twitter, Instagram, dan TikTok untuk menilai sejauh mana pola yang ditemukan bersifat spesifik platform [14][15].
3. Analisis longitudinal yang melacak pergeseran sentimen dan emosi dari waktu ke waktu, khususnya sebelum dan sesudah pengumuman kebijakan atau insiden, sebagaimana direkomendasikan Kusuma dkk. [9].
4. Ekstraksi aspek dengan pendekatan tanpa penyelia (BERTopic, LDA, KeyBERT) atau anotasi aspek terkendali mengikuti protokol SemEval [25], sehingga analisis sentimen berbasis aspek yang sesungguhnya dapat dilakukan; Pratama dkk. [8] menunjukkan penerapan BERTopic pada 19.843 komentar YouTube MBG dengan skor koherensi 0,46.
5. Klasifikasi berbantuan LLM dengan few-shot learning untuk skema emosi lengkap, disertai pengukuran kesepakatan terhadap anotasi manusia pada sampel yang lebih besar [7][16][21].
6. Validasi eksternal pada korpus independen dan oleh anotator independen, untuk menilai sejauh mana performa model dan struktur label bertahan di luar korpus ini.
7. Inferensi kausal — misalnya melalui desain interrupted time series pada peristiwa kebijakan — untuk menguji hubungan sebab-akibat yang tidak dapat dijawab oleh desain lintas-seksi penelitian ini.

Pernyataan

Kontribusi Penulis: Konseptualisasi: Harianto dan Virda Mega Ayu; Metodologi: Harianto; Perangkat Lunak: Harianto; Validasi: Harianto dan Virda Mega Ayu; Analisis formal: Virda Mega Ayu; Investigasi:

Harianto; Sumber daya: Harianto; Kurasi dan anotasi data: Virda Mega Ayu dan Harianto; Penulisan draf asli: Virda Mega Ayu; Peninjauan dan penyuntingan: Harianto dan Virda Mega Ayu; Visualisasi: Virda Mega Ayu; Supervisi: Harianto; Administrasi proyek: Harianto. Semua penulis telah membaca dan menyetujui versi final naskah.

Pendanaan: Penelitian ini dilaksanakan tanpa dukungan pendanaan eksternal dan sepenuhnya didanai secara mandiri oleh penulis.

Pernyataan Etika: Penelitian ini menggunakan data unggahan publik dan tidak melibatkan interaksi dengan subjek manusia, sehingga tidak memerlukan persetujuan komite etik. Langkah perlindungan privasi yang diterapkan diuraikan pada Bagian 3.3.

Penggunaan Kecerdasan Artifisial: Large language model digunakan sebagai instrumen pelabelan data sebagaimana diuraikan pada Bagian 3.2.2, dengan verifikasi manusia sebagaimana diuraikan pada Bagian

3.2.3. Penulis bertanggung jawab penuh atas seluruh isi naskah.

Pernyataan Ketersediaan Data: Dataset berupa data unggahan publik dari platform Threads yang telah melalui prapemrosesan, pelabelan, dan penghapusan kolom identitas. Data pendukung tersedia dari penulis korespondensi berdasarkan permintaan yang wajar, dengan mempertimbangkan batasan privasi, etika penelitian, dan kebijakan platform. Kode pipeline analisis tersedia atas permintaan.

Ucapan Terima Kasih: Penulis mengucapkan terima kasih kepada Badan Gizi Nasional dan Kementerian Kesehatan RI atas publikasi data dan informasi terkait Program MBG yang memberikan konteks kebijakan bagi penelitian ini, serta kepada para mitra bestari atas masukan yang memperbaiki naskah ini secara substansial.

Konflik Kepentingan: Penulis menyatakan tidak terdapat konflik kepentingan, baik finansial, profesional, maupun personal, yang dapat memengaruhi hasil, interpretasi, atau penyajian penelitian.

Referensi

[1] Badan Gizi Nasional, "MBG 2026 Targetkan 82,9 Juta Penerima, Fondasi Indonesia Emas," 2026. [Daring]. Tersedia: https://www.bgn.go.id/news/berita/mbg-2026-targetkan-829-juta-penerima-fondasi-indonesia-emas

[2] Kementerian Kesehatan RI, "Pemerintah Perkuat Tata Kelola Program MBG, Keselamatan Anak Jadi Prioritas," 2026. [Daring]. Tersedia: https://kemkes.go.id/id/pemerintah-perkuat-tata-kelola-program-mbg-keselamatan-anak-jadi-prioritas

[3] Universitas Gadjah Mada, "33 Ribu Pelajar Keracunan MBG, Target 3000 Porsi per SPPG Dinilai Melebihi Kapasitas," 2026. [Daring]. Tersedia: https://ugm.ac.id/id/berita/33-ribu-pelajar-keracunan-mbg-target-3000-porsi-per-sppg-dinilai-melebihi-kapasitas/

[4] A. P. Nugroho, R. Hartono, dan S. Nurmaini, "Classification of Public Opinion on the Free Nutritional Meal Program on YouTube Media Using the LSTM Method," arXiv preprint, arXiv:2604.26312, 2026.

[5] A. R. Putri, D. Haryanto, dan F. A. Bachtiar, "Public Opinion on The MBG Program: Comparative Evaluation of InSet and VADER Lexicon Labeling Using SVM on Platform X," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 2, hlm. 115–124, 2025.

[6] A. Z. Muhabbab, R. Ramadhan, dan M. Fadhil, "Sentiment Analysis of Indonesia's Free Nutritious Meal Program on Platform X (Formerly Twitter) Using IndoBERT," Jurnal ZERO, vol. 6, no. 1, hlm. 45–58, 2025.

[7] F. D. Ananda, H. Susanto, dan G. Wijaya, "Exploring Public Opinion on the 'Makan Bergizi Gratis' Program on X: A Comparative Analysis of IndoBERT-Large and NusaBERT-Large Models," Journal of Applied Informatics and Computing (JAIC), vol. 10, no. 1, hlm. 33–47, 2026.

[8] R. Pratama, B. Setiawan, dan D. P. Hostiadi, "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 3, hlm. 201–215, 2025.

[9] A. Kusuma, N. Lim, dan S. Karim, "Every Tweet Counts? Sentiment Analysis of Public Response Toward Indonesia's Free Nutritious Meal Program," dalam Proc. IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA), 2025, hlm. 112–121.

[10] R. Wirth dan J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," dalam Proc. 4th Int. Conf. on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 2000, hlm. 29–39.

[11] Y. A. Singgalen, "Comprehensive Analysis of Sentiment and Toxicity Dynamics in Tourist Vlog Reviews: A CRISP-DM Approach," Journal of Computer System and Informatics (JoSYC), vol. 5, no. 3, hlm. 512–527, 2024. doi: 10.47065/josyc.v5i3.5154

[12] F. Nugraha, A. Wibowo, dan R. Kusuma, "The Sentiment Analysis of Indonesian Startup Application Reviews Using TF-IDF+SVM and FastText: A Comparative Study," Journal of Information Technology and Computer Science (JITeCS), vol. 11, no. 1, hlm. 45–59, 2026.

[13] M. F. Kono, D. Anggraeni, dan H. Setiawan, "Comprehensive Comparison of TF-IDF and Word2Vec in Product Sentiment Classification Using Machine Learning Models," Journal of Applied Informatics and Computing (JAIC), vol. 9, no. 5, hlm. 317–328, 2026.

[14] P. Zhang, Y. He, E. U. Haq, J. He, dan G. Tyson, "The Emergence of Threads: The Birth of a New Social Network," dalam Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), 2025, hlm. 88–95.

[15] R. Khairunnas, B. P. Pagua, dan T. Arifin, "User Sentiment Dynamics in Social Media: A Comparative Analysis of X and Threads," IAES Int. Journal of Artificial Intelligence (IJ-AI), vol. 14, no. 2, hlm. 1502–1515, 2025.

[16] C. Shaw, P. LaCasse, dan L. Champagne, "Exploring Emotion Classification of Indonesian Tweets Using Large Scale Transfer Learning via IndoBERT," Social Network Analysis and Mining, vol. 15, no. 1, artikel 22, 2025. doi: 10.1007/s13278-025-01439-6

[17] A. Setiawan, R. Puspita, dan D. Hartanto, "Sentiment and Emotion Classification of Indonesian E-Commerce Reviews via Multi-Task BiLSTM and AutoML Benchmarking," arXiv preprint, arXiv:2604.24720, 2026.

[18] Badan Gizi Nasional, "Awal 2026, Program MBG Jangkau Hampir 60 Juta Penerima Manfaat," 2026. [Daring]. Tersedia: https://www.bgn.go.id/news/siaran-pers/awal-2026-program-mbg-jangkau-hampir-60-juta-penerima-manfaat

[19] J. Yushendri dan P. Musa, "Application of Natural Language Processing for Emotion Detection and Motivational Response Generation in Indonesian Text Using the CRISP-DM Method," Jurnal Sistemasi, vol. 15, no. 5, hlm. 1689–1705, 2025.

[20] Kementerian Koordinator Bidang Pangan, "Evaluasi Makan Bergizi Gratis Jateng, Menko Pangan Soroti Data dan Mutu," 2026. [Daring]. Tersedia: https://kemenkopangan.go.id/detail-berita/evaluasi-makan-bergizi-gratis-jateng-menko-pangan-soroti-data-dan-mutu

[21] F. Gilardi, M. Alizadeh, dan M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," Proceedings of the National Academy of Sciences, vol. 120, no. 30, e2305016120, 2023. doi: 10.1073/pnas.2305016120

[22] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool, 2012.

[23] P. Ekman, "An Argument for Basic Emotions," Cognition and Emotion, vol. 6, no. 3–4, hlm. 169–200, 1992. doi: 10.1080/02699939208411068

[24] R. Plutchik, "A General Psychoevolutionary Theory of Emotion," dalam Emotion: Theory, Research, and Experience, R. Plutchik dan H. Kellerman, Ed. New York: Academic Press, 1980, hlm. 3–33.

[25] M. Pontiki dkk., "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," dalam Proc. 10th Int. Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, 2016, hlm. 19–30.

[26] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, hlm. 37–46, 1960. doi: 10.1177/001316446002000104

[27] H. Cramér, Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press, 1946.

[28] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, edisi ke-2. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.