E-ISSN : 2962-9713
P-ISSN : 2963-928X

Analisis Sentimen Komentar Youtube Pada Video *"Mindfulness In The Age Of Bad News"* Oleh Maudy Ayunda Menggunakan Indobert

Miftakhul Rizki[1], Faris Syaddad Al-Muttaqin[2], Ramanda Rafi Esnawan[3], Waeisul Bismi[4], Nila Hardi[5], Imam Budiawan[6]
[1]Informatika, Universitas Bina Sarana Informatika, 15240626@bsi.ac.id
[2]Informatika, Universitas Bina Sarana Informatika, 15240600@bsi.ac.id
[3]Informatika, Universitas Bina Sarana Informatika, 15240586@bsi.ac.id
[4]Informatika, Universitas Bina Sarana Informatika, Waeisul.wbn@bsi.ac.id
[5]Informatika, Universitas Bina Sarana Informatika, nila.nad@bsi.ac.id
[6]Informatika, Universitas Bina Sarana Informatika, imam.imb@bsi.ac.id

Abstrak

Penelitian ini bertujuan mengklasifikasikan sentimen 8.360 komentar YouTube pada video "Mindfulness in the Age of Bad News" di kanal Maudy Ayunda, sekaligus menelaah ciri kebahasaan tiap kelas sentimen. Data diambil melalui YouTube Data API v3, dilabeli otomatis (pseudo-label) menggunakan IndoBERT, lalu dibagi menjadi 6.688 data latih, 836 data validasi, dan 836 data uji. Model yang dikembangkan berupa arsitektur hibrida, yakni IndoBERT (indobert-base-p1) yang dibekukan sebagai pengekstrak fitur kontekstual dan dua lapis Transformer Encoder sebagai pengklasifikasi. Teks asli ditokenisasi dengan BertTokenizer untuk pemodelan; versi yang sudah dibersihkan dan dinormalisasi dipakai untuk analisis bahasa. Ciri kebahasaan ditelaah di tingkat frasa dengan log-odds ratio. Model hibrida ini memperoleh akurasi 85,77%, presisi makro 81,28%, recall makro 83,39%, F1-score makro 82,21%, dan F1-score terbobot 85,91%, dengan waktu pelatihan 184,6 detik di GPU T4. Model ini mengungguli empat model pembanding (RoBERTa, LSTM, GRU, dan BiLSTM), dengan selisih akurasi 16,03 poin persentase terhadap model pembanding terbaik. Sentimen negatif mendominasi dataset (62,1%), dengan topik yang banyak menyentuh pajak, empati, dan kesenjangan ekonomi. Komentar positif dan netral lebih banyak ditulis dalam bahasa Inggris, dengan isi berupa dukungan yang bercampur sindiran. Sebagian kesalahan prediksi berasal dari sarkasme dan label otomatis (pseudo-label) yang kurang tepat, sehingga hasil perlu dibaca dengan mempertimbangkan keterbatasan tersebut.

Kata Kunci: *Analisis Sentimen; IndoBERT; Komentar YouTube; Maudy Ayunda; Kesehatan Mental*

Abstract

Comment sections under mental health education videos hold public responses that are far more layered than they appear. This study classifies the sentiment of 8,360 YouTube comments on the video "Mindfulness in the Age of Bad News" from Maudy Ayunda's channel and examines the linguistic features of each sentiment class. Data were collected via the YouTube Data API v3, automatically labeled (pseudo-labels) using IndoBERT, and split into 6,688 training, 836 validation, and 836 test samples. The proposed model is a hybrid architecture consisting of a frozen IndoBERT (indobert-base-p1) as the contextual feature extractor and two Transformer Encoder layers as the classifier. Linguistic features were examined at the phrase level using log-odds ratios. The hybrid model achieved an accuracy of 85.77%, a macro precision of 81.28%, a macro recall of 83.39%, a macro F1-score of 82.21%, and a weighted F1-score of 85.91%, with a training time of 184.6 seconds on a T4 GPU. It outperformed four baseline models (RoBERTa, LSTM, GRU, and BiLSTM), exceeding the best-performing baseline by 16.03 percentage points in accuracy. Negative sentiment dominates the dataset (62.1%), with topics frequently touching

E-ISSN : 2962-9713
P-ISSN : 2963-928X

on taxes, empathy, and economic inequality. Positive and neutral comments are more often written in English and contain support mixed with irony. Some prediction errors stem from sarcasm and inaccurate pseudo-labels, so the results should be interpreted with this limitation in mind.

Keywords: *Sentiment Analysis; IndoBERT; YouTube Comments; Maudy Ayunda; Mental Health*

Corresponding Author:
Faris Syaddad Al-Muttaqin
Universitas Bina Sarana Informatika; 15240600@bsi.ac.id

PENDAHULUAN

Lini masa media sosial nyaris tak pernah sepi dari berita buruk. Paparan konten negatif dilaporkan berkaitan dengan kecemasan, depresi, dan stres pada generasi muda Indonesia (Sabillillah, 2024), dan sebagian pembuat konten mulai menanggapinya dengan membahas cara menyikapi kabar buruk, salah satunya lewat video YouTube berdurasi panjang.

Video "Mindfulness in the Age of Bad News" yang dibawakan Maudy Ayunda termasuk di antaranya. Mindfulness berarti menyadari pengalaman saat ini secara penuh tanpa menghakimi dan tanpa bereaksi berlebihan (Rusdin dkk., 2025); judul videonya sudah menawarkan sikap itu sebagai pegangan di tengah arus berita buruk. Penonton merespons dengan beragam di kolom komentar: ada yang mendukung, ada yang mengkritik, ada pula yang menolak terang-terangan. Komentar YouTube pada konten edukasi kesehatan dapat mencerminkan opini publik terhadap konten tersebut (Wijaya dkk., 2026), sehingga komentar pada video ini layak dianalisis.

Opini publik di ruang daring jarang seragam. Studi sentimen berbahasa Indonesia pada isu yang memecah pendapat, misalnya isu LGBT di komentar YouTube (Oswari dkk., 2024), konflik Israel dan Iran (Saragih dkk., 2026), serta wacana politik (Sidauruk & Herowati, 2026), sama-sama berhadapan dengan opini yang beragam dan sering berseberangan. Komentar pada video kesehatan mental berpotensi serupa, sebab sebagian penonton merasa terbantu sementara yang lain skeptis. Pemetaan sentimen membuat keragaman itu terbaca lewat angka dan tidak berhenti pada kesan.

Membaca ribuan komentar satu per satu jelas tidak praktis, sehingga klasifikasi sentimen otomatis banyak dipilih sebagai jalan keluar (Ertansyah dkk., 2025). Untuk teks berbahasa Indonesia, IndoBERT, yaitu model Transformer yang telah dilatih awal pada korpus Indonesia, sering jadi andalan (Geni dkk., 2023; Marutho & Utomo, 2025; Merdiansah dkk., 2024). Komentar YouTube sendiri tidak mudah diolah. Bahasanya informal, penuh singkatan dan bahasa gaul, jadi tahap praproses sangat memengaruhi hasil (Rifaldi dkk., 2023). Sarkasme menambah masalah karena makna harfiah kalimat bisa berlawanan dengan maksud penulisnya, dan klasifikasi pun sering meleset (Riza & Akmaludin, 2026). Belum ada model yang terbukti paling unggul. Jadi membandingkan beberapa model pada data yang sama sudah lazim dilakukan (Riza dkk., 2025).

Sebagian besar penelitian sebelumnya melatih ulang seluruh bobot model. IndoBERT yang dipakai murni sebagai pengekstraksi fitur masih jarang ditemui, apalagi pada komentar YouTube bertema kesehatan, dan biasanya berpasangan dengan pengklasifikasi sederhana di luar jaringan Transformer (Wijaya dkk., 2026). Model hibrida berbasis BERT sebenarnya sudah pernah dilaporkan untuk sentimen multikelas

E-ISSN : 2962-9713
P-ISSN : 2963-928X

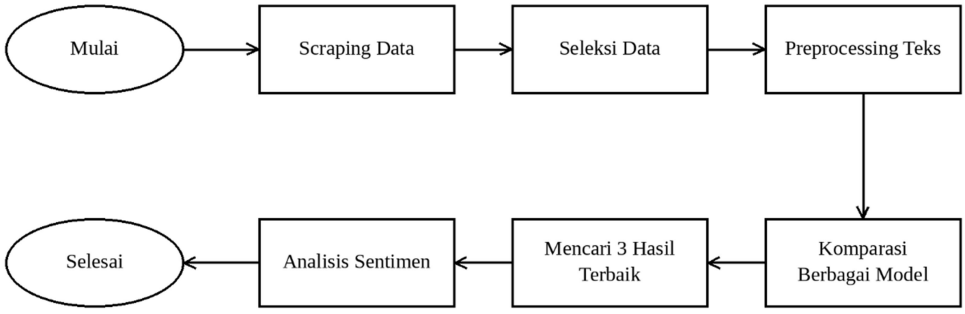
(Uyun dkk., 2025). Penelitian ini mengambil rancangan serupa dengan modifikasi: IndoBERT (indobert-base-p1) dibekukan sebagai pengekstraksi fitur kontekstual, sementara dua lapis Transformer Encoder dilatih sebagai pengklasifikasi. Hanya sekitar 1,25 juta dari 125,69 juta parameter yang ikut dilatih, dan pelatihan rampung dalam 184,6 detik di GPU T4. Ketidakseimbangan kelas ditangani dengan pembobotan kelas, bukan SMOTE, sebab data sintetis pada ruang embedding belum tentu mewakili hubungan makna yang alami (Wijaya dkk., 2026).

Di luar klasifikasi, penelitian ini membandingkan ciri bahasa tiap kelas dengan log-odds ratio pada tingkat frasa. Pilihan ini punya alasan yang cukup jelas. Frekuensi kata tunggal cenderung dikuasai kata umum yang muncul di semua kelas, dan pendekatan itu memotong frasa bermakna seperti negasi. Log-odds ratio lebih tajam karena menonjolkan unit teks yang benar-benar lebih sering hadir di satu kelas ketimbang dua kelas lainnya.

Penelitian ini bertujuan mengevaluasi model hibrida beserta beberapa model pembanding dalam mengklasifikasikan sentimen komentar pada video tersebut, sekaligus mengidentifikasi frasa khas di setiap kelas sentimen. Hasilnya diharapkan memberi gambaran yang lebih terukur tentang respons publik terhadap konten edukasi kesehatan mental.

METODE

Penelitian ini bersifat kuantitatif eksperimental dan berjalan dalam enam tahap, dari pengambilan data sampai analisis sentimen (Gambar 1). Penelitian sentimen komentar YouTube lain juga memakai alur bertahap serupa (Wijaya dkk., 2026). Tiap tahap dijelaskan berikut ini.



Gambar 1. Alur penelitian

Komentar publik pada video “Mindfulness in the Age of Bad News” diambil lewat YouTube Data API v3 pada 14 Oktober 2026, dengan cara yang sama seperti penelitian sejenis pada video bertopik kesehatan (Wijaya dkk., 2026).

Komentar kosong dan duplikat dibuang, menyisakan 8.360 komentar unik dari 7.869 akun. Setiap komentar sudah berlabel tiga kelas (negatif, netral, positif) berikut skor keyakinan label (label_score) yang berkisar dari 0,335 sampai 0,998 dengan rata-rata 0,854. Label itu bersumber dari kolom sentimen_bert, hasil klasifikasi otomatis model IndoBERT Base pralatih mdhugol/indonesia-bert-sentiment-classification yang di-fine-tune pada dataset sentimen PROSA; tidak ada anotator manusia di balik label ini. Karena sampelnya belum divalidasi secara manual, label diperlakukan sebagai pseudo-label, dan hal itu menjadi salah satu keterbatasan penelitian.

Kelas negatif mendominasi dengan 5.195 komentar (62,1%), disusul positif 1.670 (20,0%) dan netral 1.495 (17,9%). Ketimpangan begini lazim pada data media sosial (Wijaya dkk., 2026). Komentarnya cenderung pendek, dengan median 15 kata dan rata-rata 31,8 kata, walau yang terpanjang mencapai 1.102 kata. Dataset dibagi dalam dua

E-ISSN : 2962-9713
P-ISSN : 2963-928X

fase dengan stratifikasi label dan random_state 42: 80% untuk data latih, sisanya dibagi rata antara data validasi dan data uji. Eksperimen sejenis memakai rasio standar 80:20 (Wijaya dkk., 2026); di sini data validasi memantau jalannya pelatihan, sedangkan data uji baru muncul pada pelaporan akhir. Analisis hanya mencakup komentar publik, dan identitas pengguna tidak ditampilkan. Sebaran datanya ada pada Tabel 1.

Tabel 1.Distribusi kelas sentimen pada setiap subset data

Subset	Negatif	Netral	Positif	Total
Latih	4.156	1.196	1.336	6.688
Validasi	519	150	167	836
Uji	520	149	167	836
Total	**5.195**	**1.495**	**1.670**	**8.360**

Praproses menghasilkan dua versi teks, dan keduanya hanya untuk analisis bahasa. Versi ringan (text_light) cuma merapikan komentar; huruf kapital, tanda baca, dan emoji (ada pada 1.770 komentar) tetap dipertahankan karena bisa membawa emosi. Versi bersih (text_clean) melewati pembersihan URL, emoji, dan tanda baca, lalu case folding dan normalisasi kata tidak baku: yg menjadi yang, gak menjadi tidak, lu menjadi kamu, gue menjadi saya, bgt menjadi banget. Tahapan serupa pernah dipakai pada teks media sosial bertema kesehatan mental (Rifaldi dkk., 2023; Wijaya dkk., 2026).

Model sendiri menerima teks yang berbeda. BertTokenizer mendapat komentar asli dari kolom comment tanpa praproses apa pun, sebab tokenizer WordPiece IndoBERT dirancang untuk teks mentah, dan penghapusan stopword atau stemming bisa membuang konteks seperti negasi dan tanda baca. Teks dipotong sampai 500 karakter, lalu ditokenisasi dengan BertTokenizer bawaan indobert-base-p1. Tokenizer itu menambahkan token [CLS] dan [SEP], memotong dan mengisi urutan sampai 128 token, serta menyertakan attention mask.

Beberapa model dilatih pada data latih, label, dan pembagian data yang sama supaya perbandingannya adil: IndoBERT, RoBERTa, LSTM, BiLSTM, BERTopic, dan GRU. Pembandingan seperti ini lazim dalam analisis sentimen (Riza dkk., 2025). Salah satunya model hibrida yang diusulkan (Tabel 2). Bagian pertamanya, IndoBERT (indobert-base-p1), dibekukan; bobotnya tidak diperbarui dan ia hanya menghasilkan representasi kontekstual berdimensi 768 untuk tiap token, seperti pada penelitian sejenis (Wijaya dkk., 2026).

Bagian kedua adalah pengklasifikasi yang dilatih. Representasi diproyeksikan secara linear ke dimensi 256, lalu melewati dua lapis Transformer Encoder dengan dimensi model 256, jaringan umpan-maju 512, 8 head atensi, aktivasi GELU, dan dropout 0,1. Mean pooling tanpa token [CLS] meringkas keluarannya menjadi satu vektor per komentar, kemudian lapisan linear 256 memetakannya ke 3 kelas. Total parameternya 125.693.187, dan 1.251.843 di antaranya (±1,0%) ikut dilatih. Pengklasifikasi ini tidak punya embedding posisi maupun token tambahan yang bisa dipelajari. Model hibrida berbasis BERT untuk sentimen multikelas sudah pernah dilaporkan (Uyun dkk., 2025). Rincian susunan parameter ada pada Tabel 2.

Tabel 2. Konfigurasi arsitektur dan eksperimen model hibrida

Komponen	Konfigurasi
Model pralatih	indobert-base-p1, dibekukan (12 lapis, dimensi 768)
Proyeksi	Linear 768 → 256
Transformer Encoder	2 lapis; dimensi model 256; jaringan umpan-maju 512; 8 head; aktivasi GELU; dropout 0,1

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Komponen	Konfigurasi
Agregasi	Mean pooling; token [CLS] tidak dipakai sebagai masukan
Keluaran	Linear 256 → 3 kelas (negatif, netral, positif)
Total parameter	125.693.187
Parameter dilatih	1.251.843 (±1,0%)
Perangkat dan seed	GPU T4; seed 42
Jumlah epoch	10
Waktu pelatihan	184,6 detik

Pelatihan memakai cross-entropy berbobot kelas. Bobot dihitung dengan skema balanced dari data latih: 0,536 untuk negatif, 1,864 untuk netral, dan 1,669 untuk positif, sehingga kesalahan pada kelas yang lebih jarang dihukum lebih berat. Skema ini menggantikan SMOTE, yang menghasilkan data sintetis pada ruang embedding. Model hibrida dilatih 10 epoch sambil memantau loss dan F1-score makro pada data validasi (Gambar 2), dengan seed 42 di GPU T4 selama 184,6 detik. Optimizer-nya AdamW, learning rate 2e-5, ukuran batch 32. Checkpoint terbaik disimpan berdasarkan F1-score makro tertinggi pada data validasi.

Setiap model pembanding punya hyperparameter sendiri. RoBERTa memakai learning rate 2e-5 dan ukuran batch 32. Model Transformer BERT juga ber-learning rate 2e-5, tetapi batch-nya 16 untuk pelatihan dan 32 untuk evaluasi. BERTopic memakai ukuran batch 64. Ketiga model sekuensial (GRU, LSTM, BiLSTM) sama-sama dilatih dengan optimizer Adam, learning rate 1e-3, dan batch 64.

Seluruh model diuji pada 836 data uji yang tidak ikut pelatihan maupun penyetelan. Metriknya akurasi, presisi, recall, dan F1-score, semuanya dihitung dari matriks konfusi tiga kelas (Wijaya dkk., 2026). Akurasi adalah proporsi prediksi benar dari seluruh data uji. Presisi suatu kelas menunjukkan berapa banyak komentar yang diprediksi masuk kelas itu yang memang benar, sedangkan recall menunjukkan berapa banyak komentar kelas tersebut yang berhasil terdeteksi. F1-score menggabungkan keduanya sebagai rata-rata harmonik. Presisi, recall, dan F1-score dilaporkan sebagai rerata makro, yang memperlakukan ketiga kelas sama penting, dan F1-score juga dilaporkan sebagai rerata tertimbang menurut jumlah data tiap kelas. Rerata makro perlu ada karena akurasi bisa tampak tinggi pada data tidak seimbang walaupun kelas minoritas gagal dikenali (Wijaya dkk., 2026). Tiga model dengan F1-score makro tertinggi masuk ke tahap analisis lanjutan.

Tahap terakhir menganalisis sentimen dari dua sisi, yaitu sebaran kelas dan ciri bahasa tiap kelas, dengan label asli dataset. Ciri bahasa diperiksa pada tingkat kata dan frasa lewat log-odds ratio berbasis informative Dirichlet prior. Metode ini membandingkan peluang relatif sebuah kata atau frasa muncul di satu kelas dengan peluangnya di dua kelas lain, dengan prior berkekuatan 500 agar unit yang jarang muncul tidak mendominasi. Hasilnya berupa z-score; unit dengan z-score di atas 1,96 dianggap khas bagi kelasnya pada tingkat kepercayaan 95%. Pendekatan z-score log-odds dengan prior Dirichlet informatif juga pernah dipakai untuk menonjolkan kata khas pada word cloud percakapan daring (Lee dkk., 2026).

Analisis memakai text_light yang ditokenisasi ulang. Huruf diubah menjadi huruf kecil, huruf yang berulang tiga kali atau lebih diringkas jadi satu, hanya huruf alfabet yang disimpan, dan token satu huruf dibuang. Stopword formal dan partikel informal dihapus, tetapi negasi seperti tidak, bukan, belum, ga, gak, dan nggak tetap ada karena bisa mengubah makna frasa. Sebuah kata harus muncul di sedikitnya 10 komentar. Frasa

E-ISSN : 2962-9713
P-ISSN : 2963-928X

berupa n-gram dua sampai tiga kata harus muncul di sedikitnya 8 komentar, tidak boleh diawali atau diakhiri stopword, dan dihitung menurut jumlah komentar yang memuatnya. Frasa masuk jika sedikitnya 60% komentar yang memuatnya berlabel kelas bersangkutan, sementara subfrasa yang hampir selalu muncul dalam frasa lebih panjang dibuang. Keluarannya berupa word cloud berbobot z-score, 20 kata dan frasa teratas per kelas, dan sepuluh komentar berskor label tertinggi per kelas sebagai bukti kualitatif.

HASIL DAN PEMBAHASAN

Kinerja Model

Tabel 3 merangkum kinerja seluruh model pada 836 data uji. Dari lima model yang diuji, tiga dengan akurasi tertinggi adalah IndoBERT (85,77%), RoBERTa (69,74%), dan LSTM (68,06%). Jarak IndoBERT ke dua model lainnya mencolok, lebih dari 15 poin. Model berbasis Transformer tampaknya lebih piawai menangkap konteks teks informal ketimbang arsitektur sekuensial.

Tabel 3. Perbandingan kinerja seluruh model pada data uji

Model	Akurasi	Presisi Makro	Recall Makro	F1 Makro	Peringkat Akurasi
IndoBERT	85.77	81.28	83.39	82.21	1
Roberta	69.74	63.37	67.04	64.77	2
LSTM	68.06	59.32	60.49	59.30	3
GRU	66.27	59.46	62.92	60.42	4
BLSTM	65.67	60.04	64.88	61.14	5

Tabel 4 merinci metrik IndoBERT: akurasi 85,77%, F1-score makro 82,21%, dan F1-score tertimbang 85,91%. Kedua F1 berselisih sekitar 3,7 poin, pola yang lumrah pada data tidak seimbang karena kelas dominan lebih mudah dipelajari (Wijaya dkk., 2026). Presisi makro (81,28%) dan recall makro (83,39%) berdekatan, jadi model tidak condong secara sistematis ke satu arah.

Tabel 4. Hasil evaluasi model hibrida pada data uji (n = 836)

Metrik	Nilai
Akurasi	85,77%
Presisi (makro)	81,28%
Recall (makro)	83,39%
F1-score (makro)	82,21%
F1-score (tertimbang)	85,91%
Waktu pelatihan	184,6 detik (GPU T4)
Total parameter	125.693.187

Per kelas (Tabel 5), gambarnya lebih menarik. Kelas negatif paling mudah dikenali dengan F1-score 90,78%, jauh di atas netral (78,37%) dan positif (77,48%); dari 520 komentar negatif, 463 ditebak benar. Kelas netral berpola lain, karena recall-nya (83,89%) melampaui presisi (73,53%). Model jadi terlalu gampang memberi label netral: 45 komentar dari kelas lain ikut masuk, sedangkan dari 149 komentar netral hanya 24 yang terlewat. Itu sejalan dengan pembobotan kelas, yang menghukum kesalahan pada kelas minoritas lebih berat.

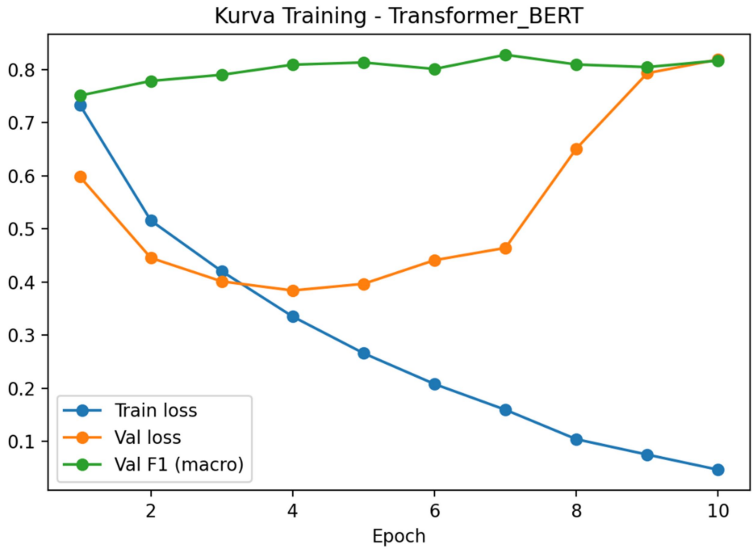
Tabel 5. Matriks konfusi dan metrik per kelas pada data uji

Aktual \ Prediksi	Negatif	Netral	Positif	Presisi (%)	*Recall* (%)	F1 (%)

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Aktual \ Prediksi	Negatif	Netral	Positif	Presisi (%)	*Recall* (%)	F1 (%)
Negatif	463	27	30	92,60	89,04	90,78
Netral	17	125	7	73,53	83,89	78,37
Positif	20	18	129	77,71	77,25	77,48

Kurva pelatihan di Gambar 2 perlu dilihat dengan teliti. Training loss turun terus dari sekitar 0,75 di epoch pertama ke sekitar 0,06 di epoch ke-10. Validation loss berjalan lain: mencapai titik terendah sekitar 0,38 pada epoch ke-4 sampai ke-5, lalu naik lagi ke sekitar 0,81 di epoch terakhir. F1 makro validasi memuncak di sekitar 0,83 pada epoch ke-7 sampai ke-8 dan setelah itu mendatar di kisaran 0,81. Gejalanya overfitting pada fungsi kerugian: model makin hafal data latih, termasuk label yang keliru, meski kemampuan generalisasinya tidak jatuh drastis. Pseudo-label otomatis yang berisik kemungkinan memperparah keadaan. Early stopping sekitar epoch ke-5 sampai ke-7 berpotensi menurunkan validation loss tanpa banyak mengorbankan F1-score, dan karena itu evaluasi akhir memakai checkpoint dengan F1 makro validasi tertinggi, bukan epoch terakhir.



Gambar 2. Kurva pelatihan model hibrida: loss pelatihan, loss validasi, dan F1-score makro validasi per epoch

Satu catatan lagi. Hanya 1.251.843 dari 125,69 juta parameter yang dilatih, dan prosesnya selesai dalam 184,6 detik di GPU T4, sehingga murah diulang. Sebagai rujukan kasar, ekstraksi fitur IndoBERT dengan pengklasifikasi sederhana pada komentar YouTube kesehatan mental dilaporkan mencapai akurasi 78% (Wijaya dkk., 2026), tetapi dataset dan skema pembagiannya berbeda sehingga perbandingan langsung tidak sah. Persoalan yang lebih mendasar: semua metrik di sini diukur terhadap pseudo-label otomatis dari model mdhugol/indonesia-bert-sentiment-classification, bukan anotasi manusia, dan kualitas label itu membatasi klaim kinerja yang boleh dibuat.

Frasa dan Kata Khas Berbasis Log-Odds Ratio

Gambar 3 dan Gambar 4 menampilkan frasa khas (dua sampai tiga kata) dan kata khas tiap kelas. Ukuran teks mengikuti z-score, bukan frekuensi mentah, dan warnanya tidak punya arti. Kata atau frasa yang umum di semua kelas sudah diredam oleh prior informatif, sedangkan frasa baru dipertahankan kalau sedikitnya 60% komentar yang

E-ISSN : 2962-9713
P-ISSN : 2963-928X

memuatnya berlabel kelas yang sama. Yang muncul di gambar karena itu bukan yang paling sering disebut, melainkan yang benar-benar membedakan satu kelas dari dua kelas lainnya.



Gambar 3. Frasa khas per kelas sentimen berbasis log-odds ratio: (a) negatif, (b) positif, (c) netral



Gambar 4. Kata khas per kelas sentimen berbasis log-odds ratio: (a) negatif, (b) positif, (c) netral

Peta tematik kelas negatif paling padat (Gambar 3a dan 4a). Satu kluster berkisar pada berita buruk: berita buruk, kabar buruk, bad news, bad reality. Kluster lain menuntut empati (punya empati, ga punya empati, minim empati) atau mengkritik soal ekonomi (orang kaya, kuliah pake, pajak rakyat, uang rakyat, duit pajak). Ada juga isu bencana dan layanan publik: kena bencana, air bersih, keracunan mbg, yang merujuk pada program Makan Bergizi Gratis. Frasa minta maaf, take down, dan mending diem menunjukkan desakan agar video ditarik atau pembicaranya angkat suara. Pada tingkat kata, kata ganti informal seperti lu, lo, kamu, dan kalian berdampingan dengan rakyat, negara, pemerintah, pajak, miskin, dan empati. Pola ini tegas: komentar negatif banyak ditujukan langsung ke figur publik dengan bingkai jarak antara yang berkecukupan dan rakyat kebanyakan (Sabillillah, 2024). Satu hal menarik, kata privilege tidak menjadi ciri kelas negatif. Gagasannya dibawakan lewat padanan lokal seperti orang kaya dan punya

E-ISSN : 2962-9713
P-ISSN : 2963-928X

empati, sementara privilege dan privileged sendiri justru muncul di kelas netral dan positif.

Kelas positif ternyata lebih rumit daripada kelihatannya (Gambar 3b dan 4b). Ada dua kelompok frasa yang sekilas berlawanan. Kelompok pertama berisi dukungan dan penerimaan: thank you for, tetap semangat, semangat kak, your peace, be nice, we can control, to protect, to care. Kelompok kedua berupa sindiran yang kebetulan berlabel positif: let them eat cake, the audacity to, the privilege of, funded by, ga pernah susah, cottage cheese. Dualitas ini penting karena kelas positif ternyata tidak seragam muatan sentimennya. Pada tingkat kata, kata fungsi bahasa Inggris seperti the, to, and, that, dan you mendominasi, disusul semangat, semoga, protect, control, healthy, dan jernih. Kata mindful dan mindfulness tampil dalam ukuran kecil, artinya daya bedanya lemah; komentar negatif pun memakai kata yang sama, hanya untuk mempertanyakan relevansi mindfulness bagi kenyataan ekonomi sebagian besar penonton.

Kelas netral paling mengejutkan (Gambar 3c dan 4c). Frasanya sedikit, tetapi muatannya berat: marie antoinette, gal gadot, queen of tone, the rich, delete this, ignore it. Analogi Marie Antoinette jelas tidak netral. Secara historis ia menempatkan pembicara sebagai elite yang abai pada rakyat, kritik yang tajam namun tanpa satu pun kata bernilai negatif secara leksikal. Kata khasnya senada: tone, deaf, empathy, rich, privilege, privileged, shame. Netral di sini tidak berarti tak berpihak; kelas ini menampung komentar berbahasa Inggris yang mengkritik secara tersirat, tanpa ungkapan sentimen eksplisit yang bisa ditangkap model pelabel.

Dominasi kata fungsi bahasa Inggris di kelas positif dan netral membuka satu keterbatasan yang harus diakui. Daftar stopword yang dipakai hanya mencakup bahasa Indonesia, sehingga komentar berbahasa Inggris menumpuk di kedua kelas itu. Estimasi kasar menunjukkan komentar yang dominan berbahasa Inggris berjumlah 7,0% di kelas negatif, 21,5% di positif, dan 42,7% di netral. Perbedaan kosakata antarkelas pun sebagian mencerminkan perbedaan bahasa, bukan semata sentimen. Polanya sekaligus menunjukkan respons publik yang terbelah cukup tegas: komentar berbahasa Indonesia berfokus pada pajak, rakyat, dan empati, sedangkan komentar berbahasa Inggris lebih sering mengkritik lewat analogi budaya seperti let them eat cake (Oswari dkk., 2024; Saragih dkk., 2026).

Analisis Kesalahan

Sebanyak 119 komentar salah diklasifikasikan, atau 14,2% dari 836 data uji (Tabel 5). Kesalahan terbanyak ada pada komentar negatif: 30 diprediksi positif dan 27 netral, sehingga 57 dari 520 komentar negatif (11,0%) tidak terdeteksi. Rerata skor keyakinan label pada prediksi yang salah hanya 0,689, jauh di bawah yang benar (0,879). Kesalahan menumpuk pada komentar yang sejak awal sudah ambigu bagi model pelabel. Contohnya ada di Tabel 6.

Tabel 6. Contoh kesalahan klasifikasi pada data uji

No.	Komentar	Label asli	Prediksi	Keterangan
1	“The irony”	Negatif	Positif	Ironi tanpa kata bernilai negatif eksplisit
2	“cara dia menyampaikan pendapat dengan menyiapkan dan menikmati makanan hahaha miris”	Negatif	Positif	Sarkasme; *menikmati* dan *hahaha* bernuansa positif
3	“Dear maudy... semoga tidurnya nyenyak dan mimpinya buruk.”	Negatif	Positif	Doa berbentuk sindiran; bagian awal

E-ISSN : 2962-9713
P-ISSN : 2963-928X

No.	Komentar	Label asli	Prediksi	Keterangan
				tampak positif
4	“Tone deaf bruh”	Negatif	Netral	Kritik implisit, sangat singkat, berbahasa Inggris
5	“Aku sebenarnya suka konten kayak gini, hanya saja tidak tepat timingnya ...”	Positif	Negatif	Sentimen campuran (apresiasi dan kritik)
6	“Privileged final boss”	Positif	Netral	Sindiran; label asli diduga keliru
7	“Banyak bersyukur mba, sini hirup asap setiap hari.”	Positif	Netral	Sarkasme tentang kabut asap; label asli diduga keliru
8	“Ini serius yg ngmng maudy ayunda?????? Hah?!!! Astaga!!!!!!”	Netral	Negatif	Ekspresi keterkejutan; polaritas ambigu

Sarkasme dan ironi adalah sumber kesalahan yang paling gampang dikenali, dan kasusnya sejalan dengan temuan pada studi klasifikasi ulasan aplikasi (Riza & Akmaludin, 2026). Komentar 1 dan 2 berlabel negatif dengan skor keyakinan 0,99, tetapi model memprediksinya positif. Kata seperti menikmati dan hahaha terbaca positif secara leksikal, padahal penulisnya bermaksud sinis, dan nada itu tidak terjangkau model. Komentar 3 lebih halus: doa yang kedengarannya baik dipasangkan dengan harapan yang justru buruk bagi orang lain. Tanpa pemahaman pragmatik, model tidak sanggup membedakan doa tulus dari doa yang ironis. Humor, ambiguitas, dan bahasa komentar YouTube yang informal sudah lama dikenal sebagai penghambat klasifikasi (Wijaya dkk., 2026).

Ada kelompok kesalahan lain yang lebih mengganggu, dan sumbernya bukan model. Komentar 6 dan 7 berlabel positif, padahal jelas terbaca sebagai sindiran; model memprediksinya netral, dan prediksi itu tampak lebih masuk akal daripada labelnya. Masalahnya ada di label. Pseudo-label otomatis membawa noise yang tidak bisa diabaikan, jadi selisih antara prediksi dan label pada kasus seperti ini hanya menunjukkan dua model mengambil keputusan berbeda, belum tentu model yang diuji yang keliru. Validasi manual pada sampel label karena itu bukan sekadar saran metodologis. Langkah itu harus dikerjakan sebelum angka kinerja dibaca sebagai ukuran kemampuan mengenali sentimen yang sebenarnya.

Kelompok ketiga berasal dari sifat datanya. Komentar 5 menyatakan suka sekaligus mengkritik waktu rilis, sehingga sentimennya campuran, dan model memprediksinya negatif. Komentar 4 terlalu singkat untuk memberi sinyal polaritas yang jelas; model memprediksinya netral. Komentar 8 hanya berisi tanda tanya dan seru yang berulang, hampir tanpa sinyal yang bisa dibaca. Komentar pendek memberi konteks yang sedikit bagi mekanisme atensi pada Transformer Encoder, dan tanpa rujukan ke isi video, ambiguitasnya nyaris mustahil diselesaikan secara otomatis.

SIMPULAN

Model hibrida yang diusulkan menunjukkan kinerja yang baik pada 836 data uji, dengan akurasi 85,77%, F1-score makro 82,21%, dan F1-score tertimbang 85,91%. Akurasinya unggul 16,03 poin persentase dari RoBERTa dan 17,71 poin persentase dari LSTM, dua model pembanding terbaik berikutnya.

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Dari sisi efisiensi, hanya 1.251.843 dari 125,69 juta parameter yang dilatih, dan pelatihan selesai dalam 184,6 detik di GPU T4. Kelas negatif paling mudah dikenali dengan F1-score 90,78%, sedangkan kelas netral (78,37%) dan positif (77,48%) lebih sulit.

Sentimen komentar timpang: kelas negatif mencakup 5.195 komentar (62,1%), disusul positif (20,0%) dan netral (17,9%). Komentar berbahasa Indonesia banyak menyinggung pajak, empati, dan kesenjangan ekonomi, sedangkan komentar berbahasa Inggris berisi campuran dukungan dan sindiran, antara lain lewat analogi let them eat cake dan Marie Antoinette. Respons publik dengan demikian terbelah menurut isi maupun bahasa.

Penelitian ini memiliki beberapa keterbatasan. Seluruh label berupa pseudo-label otomatis, bukan anotasi manusia, sehingga angka kinerja menggambarkan kesesuaian model dengan pelabel otomatis dan belum tentu kemampuan mengenali sentimen yang sebenarnya. Sebanyak 119 komentar (14,2%) salah diklasifikasikan, antara lain akibat sarkasme, ironi, sentimen campuran, komentar yang sangat pendek, dan label yang kurang tepat. Daftar stopword hanya mencakup bahasa Indonesia, sehingga perbedaan kosakata antarkelas sebagian mencerminkan perbedaan bahasa. Data juga berasal dari satu video, satu kali pengambilan, satu pembagian data, dan satu seed.

Penelitian lanjutan disarankan melakukan anotasi manual pada sampel acak berstrata oleh lebih dari satu anotator dengan pengukuran Cohen's kappa, menerapkan early stopping atau regularisasi tambahan untuk menekan overfitting, memasukkan stopword bahasa Inggris pada analisis kata khas, memperbanyak data sarkasme dan ironi serta menelaahnya dengan LIME atau metode interpretabilitas lain (Saragih dkk., 2026), dan memperluas cakupan ke beberapa video dengan rentang waktu yang lebih panjang.

Bagi kreator dan tim komunikasi figur publik, hasil ini mengisyaratkan bahwa waktu rilis dan empati terhadap audiens patut diperhatikan, terutama saat membahas kesehatan mental di tengah krisis ekonomi. Tema negatif yang dominan berkisar pada beban ekonomi, bencana, dan tuntutan empati. Bagi pengelola platform dan pendidik kesehatan mental, model yang ringan ini dapat dipakai memantau sentimen pada perangkat berkapasitas terbatas, dengan tinjauan manual untuk komentar bernada sarkasme atau berskor keyakinan rendah.

DAFTAR PUSTAKA

Ertansyah, G. P., Kusuma, R. T. C., & Sari, A. A. (2025). Analisis sentimen pada media sosial menggunakan teknik natural language processing. *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, 183–189. https://doi.org/10.47701/qgcey104

Geni, L., Yulianti, E., & Sensuse, D. I. (2023). Sentiment analysis of tweets before the 2024 elections in Indonesia using IndoBERT language models. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, *9*(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490

Lee, B., Kim, S., Menczer, F., Ahn, Y.-Y., Kwak, H., & An, J. (2026). LLMs can infer political alignment from online conversations. *arXiv preprint arXiv:2603.11253*.

Marutho, D., & Utomo, V. G. (2025). Benchmarking IndoBERT and transformer models for sentiment classification on Indonesian e-government service reviews. *Transformatika*, *23*(1), 85–95. https://doi.org/10.26623/transformatika.v23i1.12095

Merdiansah, R., Siska, S., & Ridha, A. A. (2024). Analisis sentimen pengguna X Indonesia terkait kendaraan listrik menggunakan IndoBERT. *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, *7*(1), 221–228.

E-ISSN : 2962-9713
P-ISSN : 2963-928X

Oswari, T., Murniyati, Yusnitasari, T., Nurasiah, & Wijay, S. (2024). Sentiment analysis of Indonesian YouTube reviews about lesbian, guy, bisexual and transgender (LGBT) using IndoBERT fine tuning. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, *15*(1), 26–37. https://doi.org/10.24843/LKJITI.2024.v15.i01.p03

Rifaldi, D., Fadlil, A., & Herman. (2023). Teknik preprocessing pada text mining menggunakan data tweet "Mental Health." *Decode: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Riza, F., & Akmaludin, A. A. (2026). Deteksi sarkasme pada ulasan aplikasi MyPertamina menggunakan model IndoBERT dengan pendekatan CRISP-DM. *Jurnal Ismetek*, *21*(2), 82–89.

Riza, F., Hendrakusuma, D. F., Wibowo, B., Afghani, D. Y. A., & Abdurrahman. (2025). Perbandingan kinerja algoritma klasifikasi machine learning dalam analisis sentimen ulasan mobile banking WONDR by BNI. *INTECOMS: Journal of Information Technology and Computer Science*, *8*(2), 425–436. https://doi.org/10.31539/intecoms.v8i2.14826

Rusdin, N. S., Purwasetiawatik, T. F., & Saudi, A. N. A. (2025). Gambaran mindfulness pada mahasiswa di Kota Makassar. *Jurnal Psikologi Karakter*, *5*(2), 399–405. https://doi.org/10.56326/jpk.v5i2.7473

Sabillillah, H. (2024). Analisis pengaruh paparan konten negatif di media sosial terhadap kesehatan mental Gen Z. *TUTURAN: Jurnal Ilmu Komunikasi, Sosial dan Humaniora*, *3*(1), 43–48.

Saragih, A. S., Ronaldo, D., & Saputra, A. C. (2026). Analisis sentimen publik Indonesia terhadap konflik Israel-Iran di media sosial menggunakan IndoBERT dan explainable AI (LIME). *Jurnal Teknologi Informasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika*, *20*(1), 78–87. https://doi.org/10.47111/jti.v20i1.24465

Sidauruk, V. E., & Herowati, W. (2026). IndoBERT-based sentiment analysis of political discourse on platform X: The case of Prabowo-Gibran administration. *Journal of Applied Informatics and Computing*, *10*(1), 673–683. https://doi.org/10.30871/jaic.v10i1.11586

Uyun, S., Rosalin, R. P., Sari, L. V, & Sucinta, H. H. (2025). A hybrid classification model based on BERT for multi-class sentiment analysis on Twitter. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, *11*(2), 194–205. https://doi.org/10.26555/jiteki.v11i2.30665

Wijaya, A. S., Fahmi, A., & Pramudi, Y. T. C. (2026). Sentiment classification of health education YouTube comments using IndoBERT embeddings with logistic regression and Naïve Bayes. *Journal of Applied Informatics and Computing*, *10*(3), 3005–3020.