Technologia : Jurnal Ilmiah
p-ISSN: 2086-6917, e-ISSN: 2656-8047
Homepage: https://ojs.uniska-bjm.ac.id/index.php/JIT/index
Vol. 17, No. 4, Oktober 2026

PERBANDINGAN FUNGSI *LOSS* INDOBERT UNTUK KETIDAKSEIMBANGAN KELAS PADA SENTIMEN ASPEK ULASAN HOTEL

Nur Alamsyah[1] Wagino[2] Muhammad Edya Rosadi[3]

[1,2,3]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, UNISKA MAB, Indonesia.
*Email korespondensi: nuralamsyah@uniska-bjm.ac.id

Informasi Artikel:
Dikirim: 01-09-2026; Diterima: 22-09-2026; Diterbitkan: 20-10-2026
Doi : http://dx.doi.org/10.31602/tji.v17i4.25487

Abstrak

Ulasan hotel berbahasa Indonesia menilai beberapa aspek sekaligus, tetapi sebagian besar pasangan ulasan dan aspek tidak membahas aspek tersebut sehingga distribusi kelas sangat timpang. Penelitian ini membandingkan *cross-entropy*, *weighted cross-entropy*, dan *focal loss* untuk analisis sentimen berbasis aspek pada dataset publik ulasan hotel dengan sepuluh aspek. Pembagian data resmi dipertahankan, bobot kelas dihitung dari data latih, parameter dipilih dengan data validasi, dan evaluasi dilakukan pada data uji dengan lima seed. Audit menunjukkan bahwa 80,57% pasangan berlabel netral dan label positif-negatif tidak muncul pada data uji. Pada model linier, *weighted cross-entropy* menaikkan macro-F1 dari 76,10% menjadi 79,07% secara konsisten pada kelima seed. Pada IndoBERT, macro-F1 naik dari 91,56% dengan *cross-entropy* menjadi 92,72% dengan bobot akar, 92,49% dengan bobot seimbang, dan 92,30% dengan *focal loss*, tetapi selisihnya hanya didukung uji *bootstrap* pada satu sampai tiga dari lima seed. Kenaikan terbesar terjadi pada *recall* kelas positif dan disertai penurunan presisi. Penanganan ketidakseimbangan kelas paling berguna pada model yang lemah, sedangkan pada IndoBERT manfaatnya kecil dan perlu diuji dengan beberapa seed.

Kata Kunci: *analisis sentimen berbasis aspek, focal loss, IndoBERT, ketidakseimbangan kelas, ulasan hotel*

This is an open-access article under a Creative Commons Attribution 4.0 International (CC-BY 4.0) License. Copyright ©**2026** by authors.

Pendahuluan

Ulasan daring tentang hotel jarang berisi satu penilaian saja. Dalam satu kalimat, tamu dapat memuji kebersihan kamar sekaligus mengeluhkan air panas yang tidak keluar. Analisis sentimen tingkat dokumen meringkas ulasan seperti itu menjadi satu label sehingga keluhan terhadap fasilitas tertentu tidak terlihat. Analisis sentimen berbasis aspek (*aspect-based sentiment analysis*, ABSA) menentukan polaritas untuk setiap aspek yang dibahas dalam teks [1], [2]. Bagi pengelola hotel, keluaran per aspek lebih mudah ditindaklanjuti daripada satu skor sentimen umum [3].

Untuk bahasa Indonesia, data ABSA domain hotel tersedia melalui dataset HoASA pada tolok ukur IndoNLU [4], [5]. Setiap ulasan diberi label untuk sepuluh aspek dengan empat nilai, yaitu negatif, netral, positif, dan positif-negatif. IndoBERT, model bahasa pralatih yang mengikuti arsitektur

Bidirectional Encoder Representations from Transformers (BERT) [6], menjadi salah satu pembanding utama pada tolok ukur tersebut [4]. Model ini juga dipakai dalam penelitian ABSA ulasan hotel dan pariwisata berbahasa Indonesia [7], [8], [9].

Distribusi label HoASA sangat timpang. Audit dalam penelitian ini menunjukkan bahwa 80,57% pasangan ulasan–aspek pada data latih berlabel netral, label positif hanya 4,15%, dan label positif-negatif hanya muncul sembilan kali. Pelatihan dengan *cross-entropy* biasa pada data seperti ini cenderung menghasilkan model yang condong ke kelas mayoritas, sehingga akurasi dapat tinggi walaupun kelas minoritas jarang dikenali [10], [11]. Penanganannya umumnya dilakukan dengan mengubah data melalui *resampling* atau mengubah fungsi *loss* melalui pembobotan kelas dan *focal loss* [11], [12], [13], [14].

Hasil penerapan teknik tersebut pada model bahasa Indonesia belum seragam. Pembobotan kelas pada IndoBERT untuk ulasan Google Maps menghasilkan macro-F1 sedikit di bawah model tanpa penanganan, meskipun pengenalan kelas minoritas menjadi lebih merata [15]. Pada klasifikasi judul berita, *focal loss* yang digabung dengan teknik lain menurunkan kinerja [16], sedangkan pada deteksi perundungan siber *focal loss* termasuk dalam konfigurasi terbaik [17]. Penelitian HoASA terbaru menerapkan *oversampling* acak, tetapi tidak menjelaskan apakah pembagian data resmi dipakai dan bagaimana label positif-negatif dipetakan [7]. Akibatnya, hasil tersebut sulit dibandingkan dengan tolok ukur aslinya. Dalam literatur yang ditelusuri, belum ditemukan perbandingan fungsi *loss* pada HoASA dengan protokol pelatihan dan evaluasi yang sama.

Penelitian ini membandingkan *cross-entropy* (CE), *weighted cross-entropy* (WCE), dan *focal loss* (FL) untuk ABSA pada HoASA. Pembagian data resmi dipertahankan, bobot kelas dihitung hanya dari data latih, parameter dipilih dengan data validasi, dan data uji dipakai satu kali untuk evaluasi akhir. Penelitian ini juga mengaudit dataset dan menetapkan definisi metrik secara eksplisit agar perbandingan tidak dipengaruhi oleh kelas yang tidak muncul pada data uji. Hasil yang diharapkan adalah protokol evaluasi yang dapat diulang serta bukti empiris mengenai kondisi ketika perubahan fungsi *loss* membantu kelas minoritas.

Tinjauan Pustaka

ABSA mencakup beberapa subtugas, antara lain ekstraksi istilah aspek, deteksi kategori aspek, dan klasifikasi sentimen terhadap aspek [1]. HoASA memakai formulasi kategori aspek tetap. Daftar aspek sudah ditentukan, lalu model memprediksi polaritas setiap aspek, termasuk kemungkinan aspek tersebut tidak dibahas. Survei klasifikasi sentimen berbasis aspek mencatat pergeseran dari fitur leksikal dan model berbasis mekanisme perhatian ke model bahasa pralatih [2]. Tantangan yang masih terbuka meliputi data domain khusus dan bahasa dengan sumber daya terbatas [3].

BERT melatih *encoder* Transformer dua arah dengan tugas *masked language modeling*, kemudian disesuaikan (*fine-tuning*) untuk tugas hilir [6]. IndoBERT dilatih pada korpus Indo4B dan dievaluasi pada dua belas tugas IndoNLU, termasuk HoASA [4]. HoASA berasal dari ulasan platform AiryRooms. Label positif-negatif diberikan bila satu aspek memuat sentimen berbeda untuk objek berbeda, misalnya kebersihan kasur dan kebersihan kamar mandi [4], [5]. Azhar dan Khodra mengubah tugas ini menjadi klasifikasi pasangan kalimat dengan kalimat bantu untuk BERT multibahasa pada data AiryRooms [18]. Formulasi tersebut tidak memodelkan sentimen netral, dan ulasan berpolaritas campuran dilabeli negatif. Penelitian lain melaporkan macro-F1 0,92 dengan IndoBERT dan *oversampling* acak pada HoASA [7].

Survei penanganan ketidakseimbangan kelas pada pemrosesan bahasa alami mengelompokkan metode ke dalam *resampling*, augmentasi data, modifikasi fungsi *loss*, dan strategi pelatihan bertahap [11]. Pada data multi-label, ketimpangan terjadi antarlabel dan antarnilai dalam satu label. Menggandakan satu ulasan untuk memperbanyak satu kelas juga mengubah distribusi aspek lain dalam ulasan yang sama [12]. Karena itu, modifikasi fungsi *loss* lebih mudah dikendalikan pada HoASA: data latih tidak berubah dan hanya kontribusi tiap contoh terhadap gradien yang disesuaikan. WCE memberi bobot lebih besar kepada kelas yang jarang. Cui dkk. mengusulkan bobot berbasis jumlah sampel efektif untuk menahan bobot yang terlalu besar pada kelas yang sangat kecil [13]. *Focal loss* mengurangi kontribusi contoh yang sudah mudah diklasifikasikan melalui faktor $(1 - p)^{\gamma}$ [14]. Efek teknik-teknik ini bergantung pada tingkat ketimpangan data dan metrik yang dipakai [10].

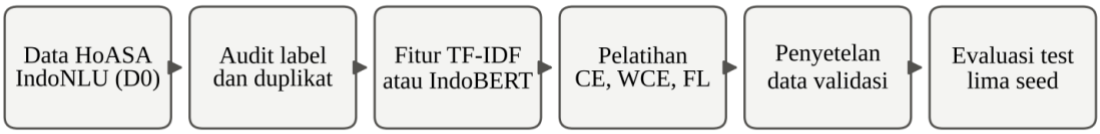
Pemilihan metrik juga memengaruhi kesimpulan. Nilai macro-F1 dapat berbeda antarimplementasi, misalnya karena cara perata-rataan atau himpunan kelas yang dimasukkan dalam perhitungan [19]. Masalah ini relevan bagi HoASA karena label positif-negatif ada pada data latih, tetapi tidak muncul pada data validasi dan data uji.

Metodologi

Jenis penelitian dan data

Penelitian ini merupakan penelitian eksperimental kuantitatif. Data diambil dari repositori IndoNLU pada folder hoasa_absa-airy dengan pembagian resmi 2.283 ulasan latih, 285 ulasan validasi, dan 286 ulasan uji. Setiap ulasan memiliki label untuk sepuluh aspek: AC, air panas, bau, umum, kebersihan, linen, layanan, sarapan, TV, dan Wi-Fi. Label uji diambil dari berkas test_preprocess.csv yang dirilis pada repositori tersebut tanggal 6 Juni 2022. Berkas test_preprocess_masked_label.csv tidak dipakai karena seluruh labelnya berisi nilai pengganti. Kesamaan ketiga berkas dengan repositori diperiksa menggunakan hash SHA-256 dan *git blob hash*.

Audit data dilakukan dengan membaca seluruh baris melalui kode. Pemeriksaan meliputi nilai label yang tidak valid, distribusi label per aspek dan per pembagian data, duplikat persis, duplikat setelah normalisasi (huruf kecil dan penghapusan tanda baca), serta duplikat hampir sama dengan kemiripan Jaccard trigram kata minimal 0,8. Data tidak dibersihkan, tidak dikoreksi, dan tidak diseimbangkan ulang. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian

Formulasi tugas dan model

Setiap ulasan diproses satu kali untuk memprediksi sepuluh label aspek secara bersamaan. Empat kelas, yaitu neg, neut, pos, dan neg_pos, tetap dipakai pada pelatihan. Dua model dipakai. Model pertama adalah baseline linier dengan fitur TF-IDF kata 1–2 gram (maksimum 10.000 fitur) dan karakter 3–5 gram (maksimum 20.000 fitur) yang muncul minimal pada dua ulasan, sehingga diperoleh 23.882 fitur. Vektorisasi dipasang hanya pada data latih, lalu diikuti satu lapisan linear dengan 40 keluaran (sepuluh aspek × empat kelas). Model kedua adalah IndoBERT indobenchmark/indobert-base-p1 dengan representasi *pooler*, *dropout* 0,1, dan sepuluh kepala klasifikasi linear yang berbagi satu *encoder*.

Fungsi loss

Fungsi *loss* dihitung pada seluruh pasangan ulasan–aspek dalam satu *batch*. Misalkan N adalah jumlah pasangan, y_i label benar pasangan ke-i, dan p_i probabilitas yang diberikan model kepada label tersebut. CE, WCE, dan FL didefinisikan pada persamaan (1)–(3).

$$L_{\mathrm{CE}} = -(1/N)\,\Sigma_i \log p_i \tag{1}$$

$$L_{\mathrm{WCE}} = -\Sigma_i\, w_{yi} \log p_i \,/\, \Sigma_i\, w_{yi} \tag{2}$$

$$L_{\mathrm{FL}} = -(1/N)\,\Sigma_i\, \alpha_{yi}\,(1 - p_i)^{\gamma} \log p_i \tag{3}$$

Bobot kelas dihitung sekali dari seluruh pasangan data latih dengan persamaan (4), dengan $K = 4$ dan n_c jumlah pasangan kelas c. Dua skema diuji, yaitu bobot seimbang yang dipotong pada nilai maksimum 10 (WCE-seimbang) dan akar kuadrat bobot seimbang (WCE-akar). Pemotongan diperlukan karena bobot seimbang kelas neg_pos mencapai 634,17 sehingga sembilan pasangan akan mendominasi gradien. Nilai batas 10 merupakan keputusan rancangan, bukan hasil penyetelan. Pada FL, $\gamma \in \{1, 2, 3\}$ dan α berupa tanpa bobot atau bobot WCE-akar. Nilai bobot ditampilkan pada Tabel 1.

$$w_c = \Sigma_k\, n_k \,/\, (K \cdot n_c) \tag{4}$$

Prosedur pelatihan dan evaluasi

Baseline linier dilatih dengan Adam, ukuran *batch* 64, maksimum 200 epoch, dan penghentian dini dengan kesabaran 15 epoch berdasarkan M1 validasi. Laju pembelajaran dipilih dari {0,01; 0,003}. IndoBERT dirancang memakai AdamW, laju pembelajaran $\{2 \times 10^{-5}; 3 \times 10^{-5}\}$, ukuran *batch* 16, *warmup* linear 10%, *weight decay* 0,01, maksimum 10 epoch, kesabaran 3 epoch, dan panjang masukan maksimum 128 token subword. Pada batas ini, lima ulasan latih (0,22%), satu ulasan validasi, dan tidak ada ulasan uji yang terpotong. Seluruh konfigurasi kedua model dirangkum pada Tabel 1. Baseline linier dijalankan pada CPU, sedangkan IndoBERT dijalankan pada GPU Tesla T4 dengan PyTorch 2.11 dan Transformers 5.20; 38 run IndoBERT memerlukan sekitar 3 jam 51 menit.

Setiap kondisi *loss* mendapat anggaran penyetelan yang sama dengan seed 42. Konfigurasi terbaik tiap kondisi dipilih berdasarkan M1 validasi, kemudian dilatih ulang dengan lima seed, yaitu 13, 21, 42, 87, dan 100. Data uji dievaluasi satu kali untuk setiap run final dan dilaporkan sebagai rerata ± simpangan baku. Selisih M1 setiap kondisi terhadap CE pada seed yang sama diuji dengan *paired bootstrap* 1.000 sampel ulang pada tingkat ulasan uji. Analisis sensitivitas mengulang evaluasi tanpa dua ulasan uji yang teksnya identik dengan ulasan latih setelah normalisasi.

Metrik evaluasi

Seluruh metrik dihitung pada pasangan ulasan–aspek yang diratakan, kecuali M3. M1 adalah macro-F1 atas kelas neg, neut, dan pos dengan daftar label eksplisit; prediksi neg_pos yang salah tetap mengurangi *recall* kelas aslinya. M1-IndoNLU adalah macro-F1 tanpa daftar label eksplisit, sama dengan fungsi evaluasi pada kode IndoNLU, dan hanya dipakai untuk perbandingan dengan literatur. M2 adalah F1 per kelas. M3 adalah rerata macro-F1 sepuluh aspek. M4 adalah micro-F1 atas pasangan (ulasan, aspek, polaritas) berlabel neg atau pos, sehingga nilainya tidak ditentukan oleh kelas netral. Akurasi dilaporkan, tetapi tidak dipakai untuk menarik kesimpulan karena sekitar 80% pasangan berlabel netral.

Tabel 1. Konfigurasi Eksperimen

Komponen	Baseline linier	IndoBERT
Representasi	TF-IDF kata 1–2 gram + karakter 3–5 gram	indobert-base-p1, *pooler*
Optimizer	Adam	AdamW, *warmup* 10%
Laju pembelajaran	{0,01; 0,003}	$\{2 \times 10^{-5}; 3 \times 10^{-5}\}$
Batch / epoch maks.	64 / 200	16 / 10
Panjang masukan maks.	–	128 token subword
Kesabaran	15 epoch	3 epoch
Bobot WCE-seimbang (neg, neut, pos, neg_pos)	1,640; 0,310; 6,027; 10	sama
Bobot WCE-akar (neg, neut, pos, neg_pos)	1,280; 0,557; 2,455; 25,183	sama
Grid FL	$\gamma \in \{1, 2, 3\}$; $\alpha \in$ {tanpa, WCE-akar}	sama
Seed penyetelan / final	42 / 13, 21, 42, 87, 100	sama

Hasil

Karakteristik data

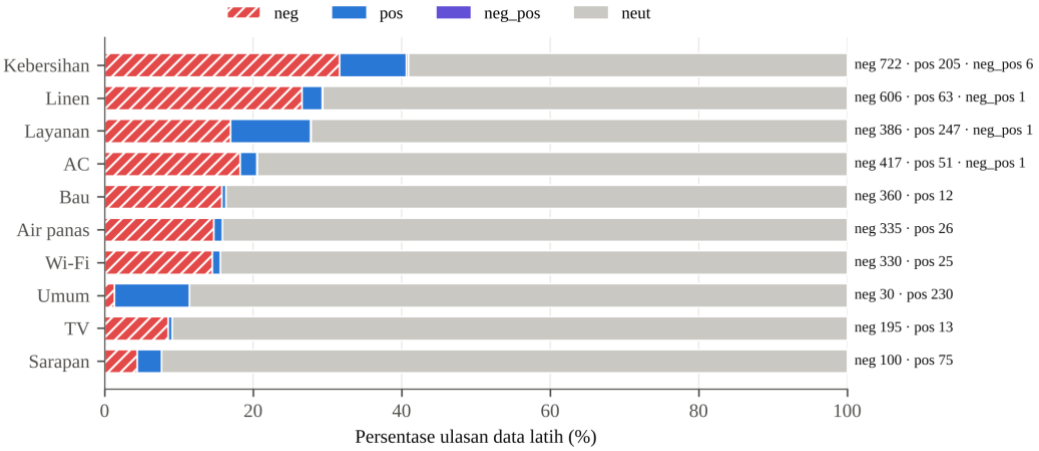
Tabel 2 menunjukkan distribusi label pada pasangan ulasan–aspek. Proporsi ketiga kelas utama hampir sama pada ketiga pembagian data, dengan kelas netral sekitar 80%. Label neg_pos hanya muncul sembilan kali pada data latih, yaitu enam pada aspek kebersihan serta masing-masing satu pada AC, linen, dan layanan. Label ini tidak muncul pada data validasi maupun data uji. Seluruh nilai label valid, tidak ada ulasan kosong, dan setiap ulasan memiliki minimal satu aspek berlabel selain netral.

Tabel 2. Distribusi Label Pasangan Ulasan–Aspek

Data	Ulasan	neg	neut	pos	neg_pos
Latih	2.283	3.481 (15,25%)	18.393 (80,57%)	947 (4,15%)	9 (0,04%)
Validasi	285	449 (15,75%)	2.273 (79,75%)	128 (4,49%)	0 (0,00%)
Uji	286	423 (14,79%)	2.298 (80,35%)	139 (4,86%)	0 (0,00%)

Persentase dihitung terhadap seluruh pasangan ulasan–aspek pada masing-masing pembagian data (jumlah ulasan × 10).

Gambar 2 memperlihatkan bahwa ketimpangan berbeda antaraspek. Kebersihan dan linen memiliki porsi negatif terbesar. Aspek umum justru didominasi label positif, sedangkan label positif pada aspek bau, TV, air panas, dan Wi-Fi hanya 12 sampai 26 ulasan. Pada aspek seperti ini, F1 kelas positif sangat peka terhadap kesalahan beberapa ulasan saja.



Gambar 2. Proporsi Label per Aspek pada Data Latih

Label neut berperan sebagai penanda bahwa aspek tidak dibahas, bukan sentimen netral dalam arti sempit. Sebagai contoh, aspek TV berlabel neut pada 2.075 ulasan latih, dan hanya 11 di antaranya memuat kata “tv”. Pola serupa terlihat pada aspek Wi-Fi, AC, dan air panas. Beberapa ulasan yang menyebut aspek secara eksplisit tetap berlabel neut, misalnya ulasan yang memuat frasa “wifi lemot”. Kasus ini menunjukkan adanya derau label. Label tidak dikoreksi agar hasil tetap sebanding dengan tolok ukur. Pemeriksaan duplikasi tidak menemukan duplikat persis maupun duplikat hampir sama lintas pembagian data. Setelah normalisasi, dua ulasan uji pendek (“kurang bersih aja” dan “AC nya kurang dingin”) identik dengan ulasan latih dan memiliki label yang sama.

Hasil penelitian terdahulu

Tabel 3 merangkum hasil yang dilaporkan pada HoASA dan data AiryRooms. Angka IndoNLU berasal dari satu seed dan dihitung dengan fungsi macro-F1 yang setara dengan M1-IndoNLU [4]. Angka penelitian lain memakai formulasi, versi data, atau prosedur penyeimbangan yang berbeda, sehingga tidak dapat dibandingkan secara langsung [18], [7].

Tabel 3. Hasil yang Dilaporkan Penelitian Terdahulu

Model	Sumber	Skor (%)	Catatan
mBERT	[4]	84,63	macro-F1, satu seed
XLM-R *base*	[4]	91,57	macro-F1, satu seed
XLM-R *large*	[4]	94,27	macro-F1, satu seed
IndoBERT *base*	[4]	92,07	macro-F1, satu seed
IndoBERT *large*	[4]	93,75	macro-F1, satu seed
IndoBERT *base* + *oversampling*	[7]	92	macro-F1; pembagian data dan pemetaan neg_pos tidak dijelaskan
mBERT + kalimat bantu	[18]	97,65	F1 biner; data AiryRooms 9.448 ulasan; netral tidak dimodelkan

Hasil baseline linier

Pada penyetelan dengan seed 42, laju pembelajaran 0,01 terpilih untuk keempat kondisi. Konfigurasi FL terbaik adalah $\gamma = 1$ dengan α akar, dengan M1 validasi 71,59%. Selisih M1 validasi antarkonfigurasi FL tidak lebih dari 1,26 poin. Hasil lima seed pada data uji ditunjukkan pada Tabel 4. Model mayoritas yang selalu memprediksi neut mencapai akurasi 80,35%, tetapi M1-nya hanya 29,70%.

Tabel 4. Hasil Baseline Linier pada Data Uji (%)

Kondisi	M1	M1-IndoNLU	F1 neg	F1 pos	M3	M4
Mayoritas (semua neut)	29,70	29,70	0,00	0,00	29,56	0,00
CE	76,10 ± 0,45	76,10 ± 0,45	75,68 ± 0,25	57,47 ± 1,16	59,88 ± 0,32	71,78 ± 0,35
WCE-seimbang	79,07 ± 0,21	79,07 ± 0,21	76,72 ± 0,47	65,56 ± 0,49	62,42 ± 0,46	74,02 ± 0,35
WCE-akar	77,22 ± 0,31	77,22 ± 0,31	76,58 ± 0,41	59,98 ± 0,66	60,56 ± 0,66	72,94 ± 0,40
FL ($\gamma = 1$, α akar)	77,55 ± 0,37	77,55 ± 0,37	76,67 ± 0,43	60,84 ± 0,83	61,02 ± 0,76	73,20 ± 0,45

Nilai berupa rerata ± simpangan baku lima seed. Tidak ada run yang memprediksi neg_pos, sehingga M1 dan M1-IndoNLU bernilai sama.

WCE-seimbang memberi nilai tertinggi pada semua metrik di Tabel 4, walaupun selisih F1 neg antara ketiga kondisi penanganan kurang dari 0,2 poin. M1 naik dari 76,10% pada CE menjadi 79,07%, dan M4 naik dari 71,78% menjadi 74,02%. Kenaikan terbesar terjadi pada kelas positif. F1 pos naik dari 57,47% menjadi 65,56% karena *recall* pos naik dari 42,59% menjadi 62,45%, walaupun presisinya turun dari 88,36% menjadi 69,00%. *Recall* kelas negatif juga naik dari 67,61% menjadi 75,18%. Sebaliknya, F1 neut turun tipis dari 95,15% menjadi 94,92%, dan akurasi turun dari 91,22% menjadi 90,92%. Gambar 3a memperlihatkan pola F1 per kelas untuk keempat kondisi.

Uji *paired bootstrap* menunjukkan bahwa selisih M1 WCE-seimbang terhadap CE memiliki interval kepercayaan 95% di atas nol pada kelima seed, dengan rerata selisih 2,97 poin. Untuk FL, interval di atas nol hanya terjadi pada 3 dari lima seed (rerata selisih 1,46 poin), sedangkan untuk WCE-akar hanya pada 1 dari lima seed (rerata selisih 1,12 poin). Perbaikan terbesar per aspek terjadi pada AC, yaitu macro-F1 naik dari 57,77% menjadi 70,51%, disusul linen (57,04% menjadi 63,96%) dan aspek umum (49,74% menjadi 55,69%). Pada kebersihan dan layanan, nilainya sedikit turun. Analisis sensitivitas tanpa dua ulasan uji yang tumpang-tindih mengubah M1 paling banyak 0,06 poin.

Hasil IndoBERT

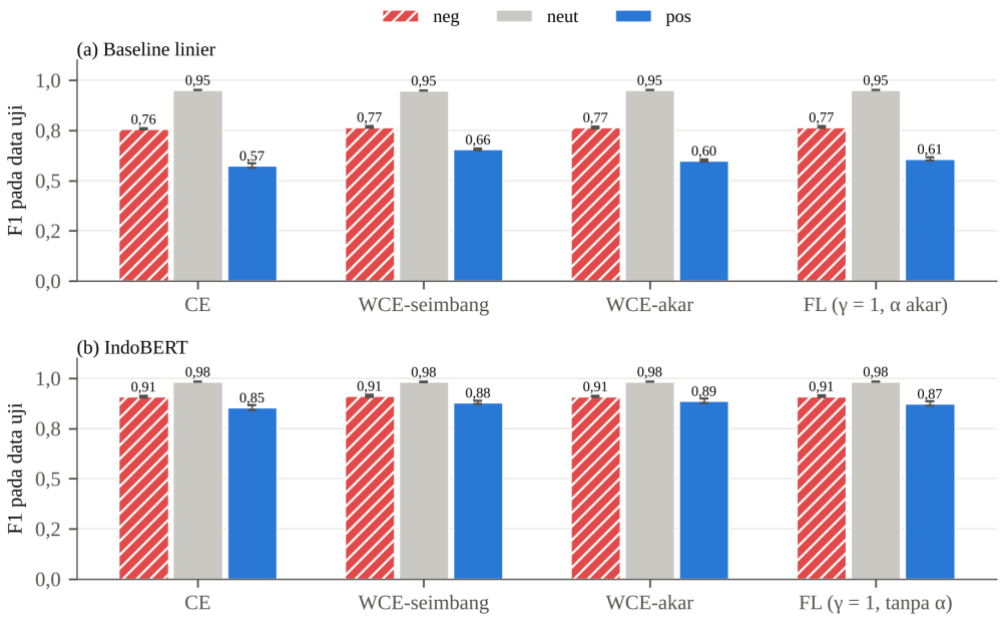
Pada penyetelan dengan seed 42, laju pembelajaran 3×10^{-5} terpilih untuk keempat kondisi. Konfigurasi FL terbaik adalah $\gamma = 1$ tanpa α, dengan M1 validasi 90,44%. Pada 10 dari 18 run penyetelan, M1 validasi terbaik dicapai pada epoch terakhir. Run penyetelan dan run final dengan seed 42 menghasilkan M1 validasi yang sama pada keempat kondisi. Hasil lima seed pada data uji ditunjukkan pada Tabel 5.

Tabel 5. Hasil IndoBERT pada Data Uji (%)

Kondisi	M1	M1-IndoNLU	F1 neg	F1 pos	M3	M4
CE	91,56 ± 0,53	91,56 ± 0,53	90,90 ± 0,46	85,47 ± 1,28	78,63 ± 1,16	89,63 ± 0,57
WCE-seimbang	92,49 ± 0,37	92,49 ± 0,37	91,28 ± 0,56	87,91 ± 0,82	86,62 ± 0,45	90,43 ± 0,48
WCE-akar	92,72 ± 0,41	92,72 ± 0,41	91,03 ± 0,32	88,74 ± 1,07	87,39 ± 0,73	90,46 ± 0,41
FL ($\gamma = 1$, tanpa α)	92,30 ± 0,57	92,30 ± 0,57	91,11 ± 0,54	87,40 ± 1,17	85,03 ± 3,37	90,21 ± 0,66

Nilai berupa rerata ± simpangan baku lima seed. Tidak ada run yang memprediksi neg_pos, sehingga M1 dan M1-IndoNLU bernilai sama.

Ketiga kondisi penanganan menghasilkan rerata M1 yang lebih tinggi daripada CE. WCE-akar memberi rerata tertinggi, yaitu 92,72% dibandingkan 91,56% pada CE, disusul WCE-seimbang (92,49%) dan FL (92,30%). Simpangan baku M1 antarseed berkisar 0,37–0,57 poin. Perubahan terbesar kembali terjadi pada kelas positif. Pada WCE-seimbang, *recall* pos naik dari 80,43% menjadi 91,08%, sedangkan presisinya turun dari 91,20% menjadi 84,98%. WCE-akar menghasilkan presisi dan *recall* pos yang lebih seimbang, yaitu 88,62% dan 88,92%. F1 neg hanya berubah 0,13–0,38 poin, F1 neut tetap sekitar 98,3%, dan akurasi seluruh kondisi berada pada rentang 96,62–96,82%. Gambar 3b memperlihatkan pola F1 per kelas tersebut.



Gambar 3. F1 per Kelas pada Data Uji: (a) Baseline Linier, (b) IndoBERT (Rerata Lima Seed)

Pada uji *paired bootstrap*, M1 WCE-akar dan FL lebih tinggi daripada CE pada kelima seed, sedangkan WCE-seimbang lebih tinggi pada empat dari lima seed. Interval kepercayaan 95% selisih M1 berada di atas nol hanya pada tiga seed untuk WCE-akar, dua seed untuk FL, dan satu seed untuk WCE-seimbang. Tidak ada kondisi dengan interval yang seluruhnya berada di bawah nol. Tidak ada run final yang memprediksi neg_pos pada data validasi maupun data uji, termasuk WCE-akar yang memberi bobot 25,18 pada kelas tersebut. Analisis sensitivitas tanpa dua ulasan uji yang tumpang-tindih mengubah M1 paling banyak 0,02 poin.

Rerata macro-F1 per aspek (M3) naik lebih besar daripada M1, dari 78,63% pada CE menjadi 87,39% pada WCE-akar dan 86,62% pada WCE-seimbang. Kenaikan ini terutama berasal dari aspek bau, TV, dan Wi-Fi. Pada WCE-akar, macro-F1 bau naik dari 61,7% menjadi 93,0%, TV dari 76,1% menjadi 96,9%, dan Wi-Fi dari 65,7% menjadi 87,9%. Data uji hanya memuat satu ulasan positif untuk bau, tiga untuk TV, dan satu untuk Wi-Fi. Sebaliknya, macro-F1 air panas turun dari 79,9% menjadi 77,6% pada WCE-akar dan 72,7% pada WCE-seimbang.

Pembahasan

Pola pada baseline linier sesuai dengan cara kerja pembobotan kelas. Bobot kelas pos pada WCE-seimbang sekitar sembilan belas kali bobot kelas neut (6,03 berbanding 0,31). Akibatnya, batas keputusan bergeser sehingga lebih banyak pasangan diprediksi pos. Pergeseran ini menaikkan *recall* kelas minoritas dengan biaya presisi, dan F1 naik karena kenaikan *recall* lebih besar daripada penurunan presisi. Pada WCE-akar, rasio bobot pos terhadap neut hanya sekitar 4,4 sehingga pergeseran batas keputusan lebih kecil. Penurunan akurasi pada saat M1 naik menunjukkan bahwa akurasi tidak cocok sebagai dasar pemilihan model untuk data ini [10], [19].

Pada IndoBERT, ruang perbaikan jauh lebih kecil. Dengan CE saja, F1 pos sudah mencapai 85,47%, sedangkan pada baseline linier hanya 57,47%. Kenaikan M1 akibat penanganan ketidakseimbangan turun dari 2,97 poin pada baseline linier menjadi 0,74–1,15 poin pada IndoBERT. Urutan kondisi juga berubah. WCE-seimbang terbaik pada baseline linier, sedangkan WCE-akar memberi rerata tertinggi pada IndoBERT. Selisih WCE-akar dan WCE-seimbang pada IndoBERT (0,23 poin) lebih kecil daripada simpangan baku antarseed, sehingga urutan keduanya belum dapat dipastikan. Pola presisi dan *recall* kelas pos menunjukkan bahwa bobot seimbang menggeser batas keputusan terlalu jauh untuk model yang sudah memisahkan sebagian besar contoh positif. *Recall* naik lebih dari sepuluh poin, tetapi presisi turun sekitar enam poin. Bobot akar menghasilkan pergeseran yang lebih kecil, sedangkan FL tanpa α hanya mengurangi pengaruh contoh yang mudah dan memberi kenaikan *recall*

pos paling kecil. Temuan ini sejalan dengan laporan bahwa teknik penanganan ketidakseimbangan pada model transformer bahasa Indonesia tidak selalu menaikkan skor keseluruhan [16], [15].

Kenaikan M3 pada IndoBERT perlu dibaca bersama jumlah contoh uji per aspek. Pada aspek bau dan Wi-Fi, data uji hanya memuat satu ulasan positif, sehingga macro-F1 aspek dapat bergeser sekitar 33 poin bergantung pada prediksi terhadap satu ulasan itu. Hal ini terlihat pada FL. Macro-F1 bau bernilai 94,9% pada seed 13, tetapi hanya 61,0% pada seed 100, sehingga simpangan baku M3 FL mencapai 3,37 poin. M1 yang dihitung pada seluruh pasangan lebih stabil sebagai metrik utama, sedangkan M3 lebih tepat dipakai sebagai diagnostik per aspek [19]. Pada baseline linier, perbaikan per aspek terutama terjadi pada AC dan linen. Aspek bau dan TV hampir tidak berubah karena kelas pos hanya memiliki 12 dan 13 contoh latih. Pembobotan global juga sedikit menurunkan nilai pada aspek dengan porsi label non-netral yang besar, seperti kebersihan dan layanan, sehingga bobot per aspek layak diuji tersendiri [12].

M1-IndoNLU IndoBERT dengan CE adalah 91,56 ± 0,53%, dengan rentang 91,02–92,38% pada lima seed. Rentang ini mencakup nilai 92,07% yang dilaporkan IndoNLU dari satu seed dengan anggaran pelatihan berbeda, yaitu maksimum 25 epoch [4]. Karena nilai satu seed dapat bergeser sekitar satu poin, perbandingan dengan Tabel 3 tidak dapat dipakai untuk menyatakan keunggulan. Rerata WCE-akar (92,72%) lebih tinggi daripada angka IndoBERT *base* tersebut, tetapi masih di bawah IndoBERT *large* (93,75%) dan XLM-R *large* (94,27%). Bobot neg_pos sebesar 10 pada WCE-seimbang dan 25,18 pada WCE-akar tidak membuat model memprediksi kelas tersebut, sehingga M1 dan M1-IndoNLU tetap sama. Pada kedua model, M1 validasi lebih rendah daripada data uji, yaitu 5,06–6,23 poin pada baseline linier dan 1,78–2,37 poin pada IndoBERT. Perbedaan ini menegaskan pentingnya beberapa seed dan larangan memakai data uji untuk memilih konfigurasi.

Penelitian ini memiliki beberapa keterbatasan. Pertama, kelas neg_pos tidak dapat dievaluasi karena tidak muncul pada data validasi dan data uji, sehingga manfaat bobot besar untuk kelas tersebut tidak dapat diukur. Kedua, label neut menggabungkan aspek yang tidak dibahas dengan sentimen netral, dan sebagian label tampak tidak sesuai dengan isi ulasan. Ketiga, hanya satu dataset dan satu domain yang dipakai. Keempat, nilai batas bobot 10 ditetapkan tanpa penyetelan karena data validasi tidak memuat neg_pos. Kelima, 12 dari 20 run final IndoBERT mencapai M1 validasi terbaik pada epoch terakhir, dan laju pembelajaran terbesar dalam grid terpilih untuk semua kondisi. Model mungkin belum konvergen penuh, sehingga hasil IndoBERT berlaku untuk anggaran pelatihan sepuluh epoch. Keenam, sebagian operasi GPU tidak sepenuhnya deterministik, walaupun run dengan seed 42 menghasilkan nilai validasi yang sama pada tahap penyetelan dan tahap final.

Kesimpulan

Audit menunjukkan bahwa label netral pada HoASA lebih tepat dibaca sebagai penanda aspek yang tidak dibahas, dan label positif-negatif tidak dapat dievaluasi dengan pembagian data resmi. Kedua hal ini perlu dinyatakan dalam setiap laporan hasil HoASA, bersama definisi metrik yang menyebutkan kelas yang dihitung. Manfaat pembobotan kelas jauh lebih besar pada baseline linier daripada pada IndoBERT. Pada baseline linier, bobot seimbang dengan batas menaikkan M1 sebesar 2,97 poin secara konsisten pada kelima seed. Pada IndoBERT, ketiga metode menaikkan rerata M1 sekitar satu poin, terutama melalui *recall* kelas positif, tetapi selisihnya terhadap CE hanya didukung uji *bootstrap* pada sebagian seed. Kenaikan macro-F1 per aspek yang besar pada IndoBERT berasal dari aspek dengan satu sampai tiga contoh positif pada data uji, sehingga tidak dapat dijadikan dasar kesimpulan.

Penelitian berikutnya disarankan memperpanjang anggaran pelatihan IndoBERT karena sebagian besar run mencapai nilai terbaik pada epoch terakhir, menguji bobot berbasis jumlah sampel efektif dan bobot per aspek, memisahkan deteksi aspek dari klasifikasi polaritas, serta menyusun data evaluasi yang memuat lebih banyak contoh positif dan positif-negatif.

Daftar Pustaka

[1] W. Zhang, X. Li, Y. Deng, L. Bing, dan W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11019–11038, 2023, doi: 10.1109/TKDE.2022.3230975.

[2] G. Brauwers dan F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Computing Surveys*,

vol. 55, no. 4, Art. no. 65, 2023, doi: 10.1145/3503044.
[3] A. Nazir, Y. Rao, L. Wu, dan L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 845–863, 2022, doi: 10.1109/TAFFC.2020.2970399.
[4] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, dan A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," dalam *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, 2020, pp. 843–857, doi: 10.18653/v1/2020.aacl-main.85.
[5] A. N. Azhar, M. L. Khodra, dan A. P. Sutiono, "Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting," dalam *Proceedings of the 2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019, pp. 35–40.
[6] J. Devlin, M.-W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," dalam *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[7] T. Aprilah, D. R. I. M. Setiadi, dan W. Herowati, "Enhancing aspect-based sentiment analysis via Hugging Face fine-tuned IndoBERT," *Journal of Applied Informatics and Computing*, vol. 9, no. 6, pp. 3821–3830, 2025, doi: 10.30871/jaic.v9i6.11409.
[8] S. Apriliani, A. Erfina, dan C. Warman, "Fine-tuned IndoBERT for aspect-based sentiment analysis of Indonesian five-star hotel reviews," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 4, pp. 437–445, 2025, doi: 10.32736/sisfokom.v14i4.2491.
[9] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, dan R. P. Agusdin, "Domain-specific fine-tuning of IndoBERT for aspect-based sentiment analysis in Indonesian travel user-generated content," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, 2025, doi: 10.20473/jisebi.11.1.30-40.
[10] J. M. Johnson dan T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019, doi: 10.1186/s40537-019-0192-5.
[11] S. Henning, W. Beluch, A. Fraser, dan A. Friedrich, "A survey of methods for addressing class imbalance in deep-learning based natural language processing," dalam *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, 2023, pp. 523–540, doi: 10.18653/v1/2023.eacl-main.38.
[12] A. N. Tarekegn, M. Giacobini, dan K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, Art. no. 107965, 2021, doi: 10.1016/j.patcog.2021.107965.
[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, dan S. Belongie, "Class-balanced loss based on effective number of samples," dalam *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277.
[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, dan P. Dollár, "Focal loss for dense object detection," dalam *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
[15] R. Widayanti dan F. C. Kasih, "Comparative analysis of baseline IndoBERT, class-weighted IndoBERT, and SMOTE with support vector machine for handling imbalanced sentiment classification in Indonesian," *Jurnal Teknik Informatika (JUTIF)*, vol. 7, no. 3, pp. 2857–2875, 2026, doi: 10.52436/1.jutif.2026.7.3.5692.
[16] E. Sirait dan J. Ismail, "Comparative evaluation of IndoBERT-based architectures for imbalanced Indonesian news title classification," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 10, no. 3, pp. 1896–1907, 2026, doi: 10.33395/sinkron.v10i3.16209.
[17] F. U. Nafiah, A. T. Damaliana, dan K. M. Hindrayani, "Indonesian cyberbullying detection using IndoBERTweet-BiGRU model on class-imbalanced X (Twitter) data," *Journal of Applied Informatics and Computing*, vol. 10, no. 4, pp. 3837-3848, 2026, doi: 10.30871/jaic.v10i4.13686.
[18] A. N. Azhar dan M. L. Khodra, "Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis," arXiv:2103.03732, 2021.
[19] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.