SEMINAR NASIONAL 2026 Universitas Ma Chung
Prosiding Seminar Nasional Universitas Ma Chung 2026
(Sistem Informasi dan Informatika)
pp. 43-57
E-ISSN: 2808-7771

ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI RUANGGURU BERBASIS INDOBERT UNTUK EVALUASI KUALITAS LAYANAN PENDIDIKAN DIGITAL

Nasywa Salsabiila Romadhona[1], Dwi Wahyu Prabowo[2], Nurahman[3]

[1]Program Studi Sistem Informasi, Universitas Darwan Ali, Jl. Batu Berlian No. 10, Kotawaringin Timur, Kalimantan Tengah, Indonesia, 74323

***Email korespondensi:** nasywasalsabiila03@gmail.com

Received: 27 05 2026 – Revised: 29 07 2026 - Accepted: 04 08 2026 - Published: 05 08 2026

Abstrak. Pertumbuhan pesat platform pendidikan teknologi (*edtech*) di Indonesia mendorong kebutuhan akan mekanisme evaluasi kualitas layanan yang lebih mendalam dari sekadar agregat rating bintang. Penelitian ini bertujuan mengembangkan sistem analisis sentimen berbasis IndoBERT untuk mengklasifikasikan ulasan pengguna aplikasi Ruangguru dari Google Play Store ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif, guna mendukung evaluasi kualitas layanan pendidikan digital secara berbasis teks. Data dilabeli secara otomatis berdasarkan rating pengguna: positif (rating 4–5), netral (rating 3), dan negatif (rating 1–2). Model IndoBERT-base-p1 (*indobenchmark/indobert-base-p1*) digunakan sebagai model dasar yang kemudian di-*fine-tuning* dengan empat skenario penanganan ketidakseimbangan kelas, yaitu tanpa penyeimbangan, *random oversampling*, *cost-sensitive learning* (*weighted cross-entropy*), dan *focal loss* (γ=2,0). Validasi dilakukan menggunakan 10-*Fold Stratified Cross-Validation* untuk memastikan kestabilan dan representativitas hasil, serta uji statistik Wilcoxon *Signed-Rank Test* untuk membandingkan performa antar skenario secara signifikan. Hasil terbaik diperoleh pada skenario tanpa penyeimbangan dengan F1-*score* sebesar 0,6951. Penelitian ini menunjukkan bahwa IndoBERT mampu menangkap nuansa bahasa Indonesia informal dalam ulasan pengguna aplikasi edtech dan berpotensi menjadi fondasi sistem evaluasi layanan pendidikan digital yang lebih akurat dan berbasis data opini nyata.

Kata kunci: Analisis Sentimen, IndoBERT, Pembelajaran Daring, *Transfer Learning*, Ulasan Pengguna.

PENDAHULUAN

Sektor pendidikan berbasis teknologi (*educational technology atau edtech*) di Indonesia mengalami pertumbuhan yang signifikan dalam satu dekade terakhir, dipercepat oleh pandemi COVID-19 yang memaksa peralihan masif dari pembelajaran tatap muka ke platform digital (Nasution, 2022).Aplikasi pembelajaran daring seperti Ruangguru, Zenius, dan Quipper telah digunakan oleh jutaan pelajar di seluruh nusantara, menjadikan ulasan pengguna di platform distribusi aplikasi seperti Google Play Store sebagai sumber umpan balik yang sangat berharga. Ulasan tersebut memuat pengalaman autentik pengguna yang mencerminkan persepsi terhadap kualitas konten, kemudahan antarmuka, ketepatan fitur, dan responsivitas layanan; dimensi-dimensi yang tidak selalu tertangkap oleh rating numerik semata (Kyriakidis & Tsafarakis, 2025).

Analisis sentimen secara otomatis terhadap ulasan berbahasa Indonesia menghadapi tantangan yang tidak trivial. Teks ulasan ditulis dalam bahasa sehari-hari yang kaya akan singkatan, campur kode (*code-mixing*) antara bahasa Indonesia, bahasa daerah, dan bahasa

Inggris, serta ekspresi informal yang sulit ditangani oleh metode berbasis leksikon atau machine learning klasik seperti *Naïve Bayes* dan SVM (Prasetiadi et al., 2024). Model berbasis transformer, khususnya BERT (*Bidirectional Encoder Representations from Transformers*), telah terbukti melampaui metode klasik dalam tugas klasifikasi teks karena kemampuannya menangkap konteks dua arah secara mendalam (Mao et al., 2023). IndoBERT, sebagai varian BERT yang dilatih khusus pada korpus bahasa Indonesia berskala besar, memberikan representasi linguistik yang lebih relevan untuk teks berbahasa Indonesia dibanding model multilingual generik (Dwi Purnomo & Sutopo, 2024).

Salah satu tantangan krusial dalam analisis sentimen pada dataset ulasan aplikasi nyata adalah ketidakseimbangan distribusi kelas (*class imbalance*). Ulasan pada platform seperti Google Play Store cenderung didominasi oleh kelas sentimen positif, sementara ulasan netral jauh lebih jarang ditemukan. Kondisi ini menyebabkan model yang dilatih tanpa penanganan khusus cenderung bias terhadap kelas mayoritas dan gagal mendeteksi kelas minoritas secara akurat (Taskiran et al., 2025). Berbagai strategi telah diusulkan untuk mengatasi permasalahan ini, di antaranya oversampling pada level data, pembobotan loss function melalui cost-sensitive learning, dan focal loss yang secara adaptif memberikan bobot lebih besar pada sampel yang sulit diklasifikasikan (Cao Luand Liu, 2022) Namun, studi komparatif yang secara sistematis mengevaluasi strategi-strategi tersebut pada konteks IndoBERT untuk teks edtech Indonesia masih sangat terbatas.

Penelitian ini bertujuan membandingkan empat skenario penanganan *class imbalance* pada model IndoBERT-*base-p1* dalam analisis sentimen ulasan pengguna aplikasi Ruangguru, guna mengidentifikasi pendekatan terbaik untuk mendukung evaluasi kualitas layanan pendidikan digital berbasis opini pengguna.

MASALAH

Penelitian ini mengidentifikasi tiga permasalahan utama. Pertama, dominasi ulasan positif pada dataset Google Play Store *class imbalance* yang berdampak pada bias model terhadap kelas mayorita, sehingga deteksi kelas minoritas, khususnya sentimen netral, menjadi rendah dan tidak representative (Kaope & Pristyanto, 2023). Kedua, belum tersedia kajian komparatif yang secara sistematis dan terukur membandingkan strategi penanganan *class imbalance*, meliputi *random oversampling*, *cost-sensitive learning*, dan *focal loss*, dalam konteks *fine-tuning* IndoBERT pada dataset ulasan aplikasi *edtech* berbahasa Indonesia. Ketiga, evaluasi kualitas layanan *edtech* yang selama ini bertumpu pada agregasi

rating numerik tidak mampu mengungkap aspek kualitatif yang sesungguhnya dikeluhkan atau diapresiasi oleh pengguna, sehingga pendekatan analisis teks mendalam diperlukan untuk *actionable insight* bagi perbaikan layanan secara berkelanjutan.

METODE PELAKSANAAN

Pengumpulan dan Pelabelan Data

Data penelitian bersumber dari dataset publik ulasan pengguna aplikasi Ruangguru pada platform Google Play Store yang tersedia melalui Kaggle. Dataset ini terdiri dari dua berkas ulasan yang digabungkan sebelum diproses lebih lanjut. Pelabelan sentimen dilakukan secara otomatis berdasarkan rating yang diberikan pengguna, dengan ketentuan sebagai berikut: ulasan dengan rating 4–5 bintang dikategorikan sebagai sentimen positif, rating 3 bintang sebagai sentimen netral, dan rating 1–2 bintang sebagai sentimen negatif. Skema pelabelan ini mengacu pada konvensi yang umum digunakan dalam penelitian analisis sentimen berbasis ulasan platform digital (Setyo Nugroho et al.).

Preprocessing

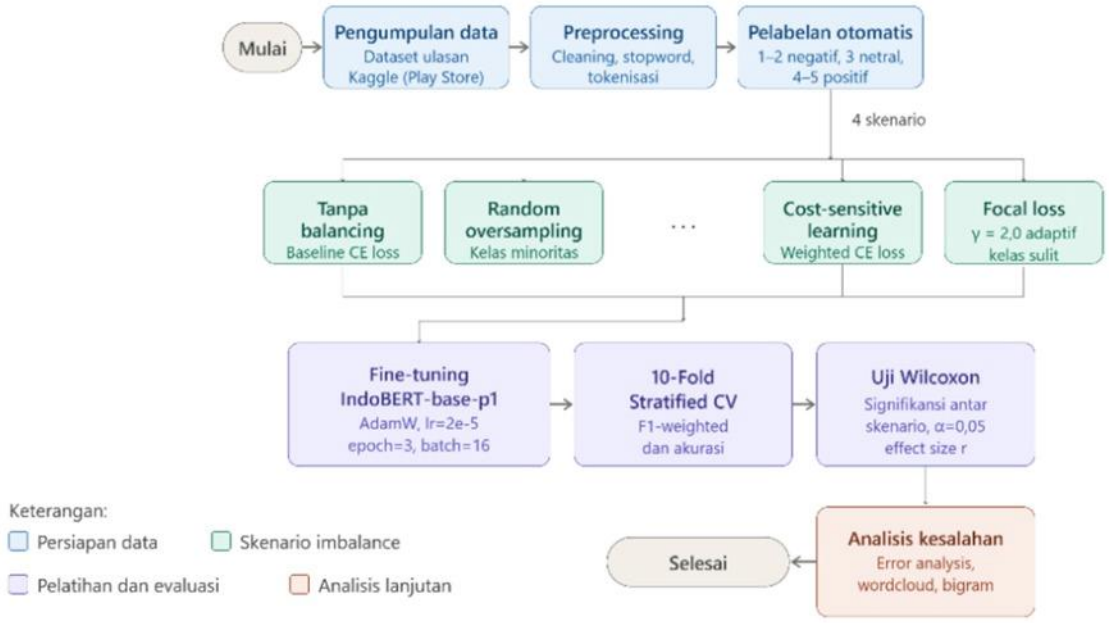
Seluruh teks ulasan melalui tahap praproses sebelum diumpankan ke model. Tahapan praproses meliputi konversi teks ke huruf kecil (*lowercase*), penghapusan URL dan karakter non-alfanumerik, penghapusan stopword berbahasa Indonesia menggunakan pustaka Sastrawi, serta tokenisasi menggunakan tokenizer bawaan IndoBERT dengan panjang sekuens maksimal (*max_len*) sebesar 128 token. Pemotongan pada batas 128 token dipilih untuk menyeimbangkan antara cakupan informasi teks ulasan dan efisiensi komputasi selama pelatihan (Jaiswal & Milios, 2023).

Penanganan Class Imbalance

Empat skenario penanganan ketidakseimbangan kelas diujicobakan secara terpisah. Skenario pertama (*Baseline*) melatih model tanpa modifikasi distribusi data maupun fungsi loss, sehingga berfungsi sebagai pembanding. Skenario kedua menerapkan random oversampling pada kelas minoritas hingga distribusi kelas menjadi seimbang sebelum pelatihan. Skenario ketiga menggunakan cost-sensitive learning dengan *weighted cross-entropy loss*, di mana bobot setiap kelas ditentukan secara invers proporsional terhadap frekuensinya dalam data latih. Skenario keempat menerapkan focal loss dengan parameter γ=2,0 yang secara adaptif mereduksi kontribusi sampel yang mudah diklasifikasikan sehingga model lebih berfokus pada sampel yang sulit (Cao Luand Liu, 2022).

Alur Penelitian

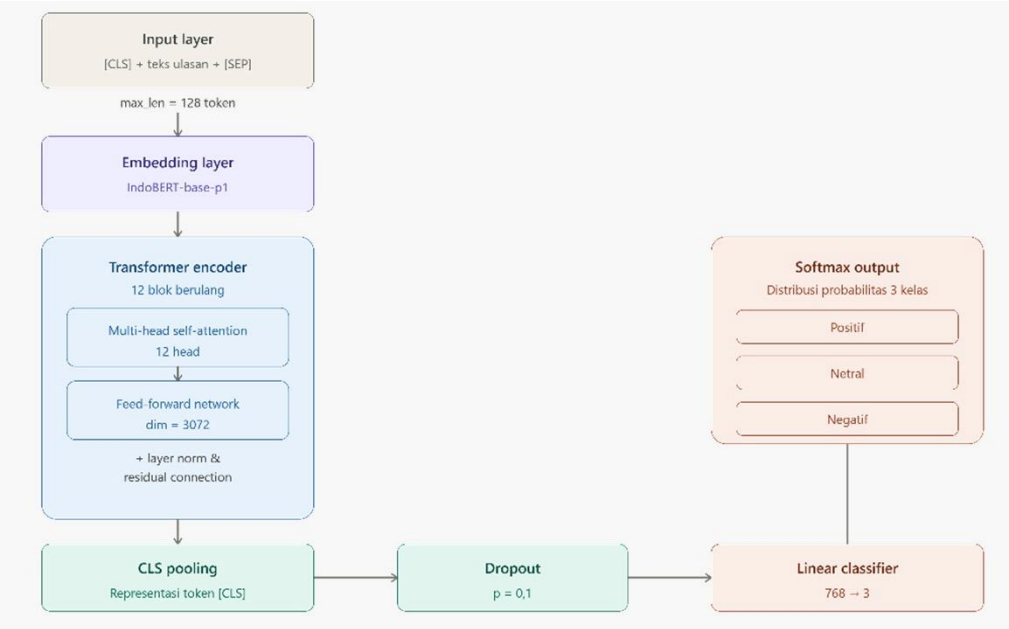
Gambar 1 menyajikan alur keseluruhan penelitian yang mencakup delapan tahap, yaitu pengumpulan data, *preprocessing*, pelabelan otomatis, pembagian empat skenario, *fine-tuning* IndoBERT, 10-*fold* CV, uji Wilcoxon, serta analisis kesalahan dan kesimpulan. Setiap tahap dirancang berurutan, dengan tahap pembagian skenario bercabang menjadi empat jalur paralel yang kembali menyatu pada tahap validasi silang sebelum berlanjut ke pengujian statistik dan interpretasi akhir.



Gambar 1. *Flowchart alur penelitian indobert*

Arsitektur dan Model Matematis

Model yang digunakan adalah IndoBERT-base-p1 (*indobenchmark/indobert-base-p1*) yang di-*fine-tuning* untuk tugas klasifikasi tiga kelas dengan menambahkan lapisan klasifikasi linier di atas representasi token [CLS]. Arsitektur model ditampilkan pada Gambar 2.



Gambar 2. *Arsitektur indobert klasifikasi sentimen*

Teks ulasan ditokenisasi dengan panjang maksimal *max_len*=128, lalu diproses melalui 12 blok *transformer encoder* yang masing-masing terdiri dari 12 *head self-attention*, FFN berdimensi 3072, *layer normalization*, dan koneksi residual.Representasi vektor [CLS] dari lapisan terakhir diproses melalui *dropout* (p=0,1) dan *linear classifier* (768 ke 3 kelas), kemudian dinormalisasi dengan fungsi *softmax* pada persamaan 1.

$$P(y_k|x) = \frac{\exp(z_k)}{\Sigma_j \exp(z_j)}, \quad (1)$$

Pelatihan menggunakan optimizer AdamW dengan *learning rate* 2×10^{-5}, 3 epoch, dan batch 16, disertai *linear warmup* pada 10% langkah pertama untuk menstabilkan pembaruan gradien awal (Antariksa et al., 2025). Fungsi loss yang digunakan pada masing-masing skenario adalah *standard cross-entropy* pada persamaan 2,

$$L = -\Sigma_c\, y_c \log(p_c), \quad (2)$$

Weighted Cross-Entropy Loss untuk *cost-sensitive learning* pada persamaan 3 (Cao Luand Liu, 2022).

$$L\_w = -\Sigma_c\, w_c \cdot y_c \cdot \log(p_c), \quad (3)$$

dan *focal loss* dengan γ=2,0 pada persamaan 4.

$$FL(p_t) = -\alpha_t(1-p_t)\text{^}\gamma \log(p_t), \text{dengan } \gamma = 2{,}0, \quad (4)$$

Evaluasi performa model menggunakan F1-*score* berbobot, yang dirumuskan pada persamaan 5.

$$F1 = \Sigma_c \left(\frac{n_c}{N}\right) \cdot \frac{2P_c R_c}{(P_c + R_c)}, \quad (5)$$

dan *effect sice Wilocoxon* pada persamaan 6 (Tapio, 2025).

$$r = \frac{Z}{\sqrt{N}}, \quad (6)$$

Evaluasi

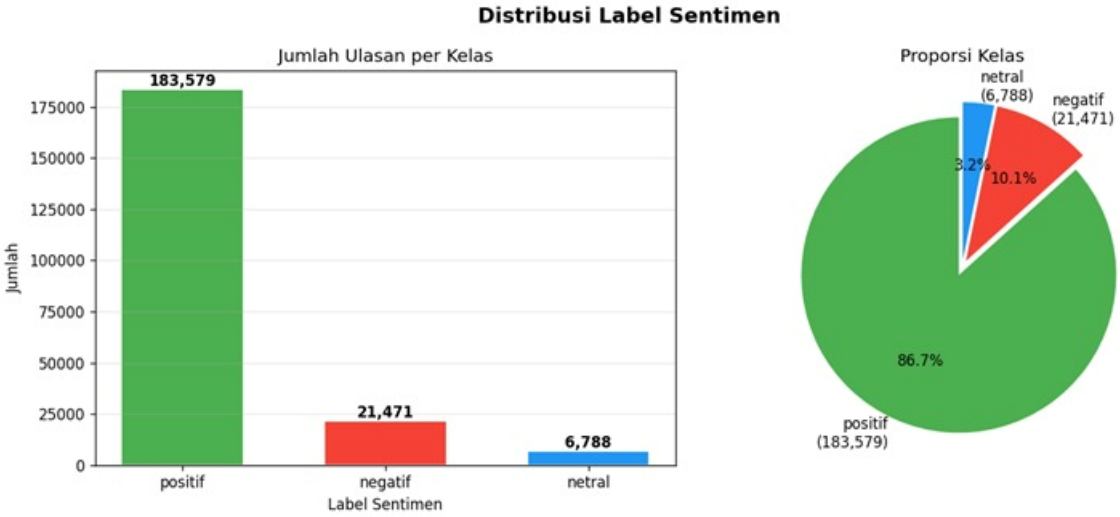
Evaluasi model dilakukan menggunakan kerangka 10-*Fold Stratified Cross-Validation*, yang memastikan proporsi kelas pada setiap lipatan (*fold*) mencerminkan distribusi kelas keseluruhan dataset. Metrik utama yang digunakan adalah F1-*score* berbobot (*weighted F1*) dan akurasi. Untuk menguji signifikansi statistik perbedaan performa antar skenario, diterapkan uji *Wilcoxon Signed-Rank Test* pada taraf signifikansi α=0,05 dengan kalkulasi *effect size r* untuk mengukur besaran perbedaan yang bermakna secara praktis (Tapio, 2025).

HASIL DAN PEMBAHASAN

Karakteristik Dataset

Dataset dibangun dari dua berkas *scrape* ulasan Play Store Ruang Guru yang digabung, menghasilkan 211.838 ulasan mentah. Setelah pembersihan menggunakan ekspresi reguler, penghapusan stopword berbasis Sastrawi, dan deduplikasi, tersisa 209.343 ulasan valid. Label sentimen diturunkan langsung dari skor rating: 1–2 sebagai negatif, 3 sebagai netral, dan 4–5 sebagai positif.

Gambar 3 memperlihatkan distribusi label dataset dan perbandingan proporsi antar kelas secara visual, yang dengan jelas menggambarkan ketimpangan yang sangat besar antara kelas positif dan kelas minoritas. Sebagaimana ditunjukkan pada Gambar 3, kelas positif mendominasi dataset dengan 183.579 ulasan (86,70%), diikuti kelas negatif sebanyak 21.471 ulasan (10,10%), dan kelas netral sebagai kelas paling minoritas dengan 6.788 ulasan (3,20%) dari total 209.343 ulasan.



***Gambar 3**. Distribusi label dataset dan perbandingan proporsi kelas*

Distribusi ini mencerminkan pola umum ulasan aplikasi mobile: pengguna puas cenderung memberi bintang tinggi tanpa narasi panjang, sementara pengguna kecewa lebih termotivasi menulis. *Imbalance Ratio* sebesar 27,04 antara kelas positif dan netral menjadikan dataset ini representatif namun juga menantang. Dominansi kelas mayoritas berpotensi mendistorsi evaluasi model secara sistematis jika tidak ditangani di level desain eksperimen. Untuk eksperimen utama, diambil sampel stratifikasi seimbang 3.000 ulasan per kelas (total 9.000, *random_state*=42).

Eksperimen Awal pada Distribusi Alami

Sebelum eksperimen utama, dilakukan uji awal pada subsampel 15.000 ulasan dengan distribusi alami yang tetap timpang (positif 12.772; negatif 1.672; netral 556), menggunakan 10-*fold CV*, dua *epoch*, dan *max_len*=64.

Tabel 1. Performa Model pada Eksperimen Awal (Distribusi Alami)

Skema	**F1-*macro***	**Akurasi**
Standard cross-entropy	0,5698	0,8967
Cost-sensitive (weighted CE)	0,5897	0,8668

Hasil pada Tabel 1 memperlihatkan *accuracy paradox* yang klasik, yaitu model *standard cross-entropy* mencatat akurasi 89,67% namun F1-macro hanya 0,5698 karena model cenderung mendominasi prediksi kelas positif. Penerapan bobot kelas pada skema cost-sensitive menaikkan F1-*macro* ke 0,5897 dengan konsekuensi penurunan akurasi ke 86,69%, sementara recall kelas netral pada *split final* hanya mencapai 0,43. Temuan ini

memotivasi perubahan desain pada eksperimen utama berupa penyeimbangan stratifikasi antar kelas dan penggunaan F1-*weighted* sebagai acuan evaluasi primer.

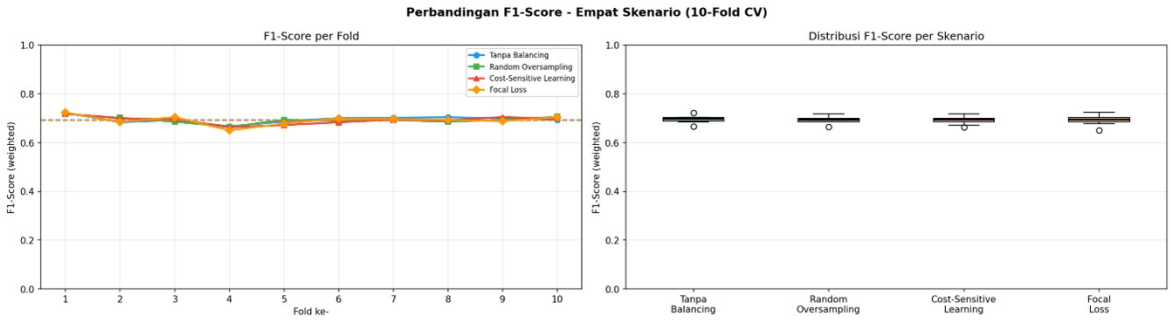
Eksperimen Utama: Perbandingan Empat Skenario Penyeimbangan Kelas

Eksperimen utama dijalankan pada sampel seimbang 9.000 ulasan dengan konfigurasi yang lebih matang: 10-*fold CV*, tiga *epoch, max_len*=128, *batch*=16, learning rate 2e-5. Empat skenario dibandingkan secara sistematis. Tabel 2 merangkum hasil performa keempat skenario berdasarkan F1-*weighted* rata-rata dan simpangan baku di sepanjang 10 *fold*.

Tabel 2. Performa Empat Skenario pada Sampel Seimbang (*F1-weighted, 10-fold CV*)

Skenario	*F1-weighted* ± SD	Akurasi
(1) Tanpa *balancing*	0,6951 ± 0,0138	0,6970
(2) *Random oversampling*	0,6927 ± 0,0134	0,6951
(3) *Cost-sensitive learning*	0,6920 ± 0,0152	0,6943
(4) *Focal loss* (γ=2,0)	0,6922 ± 0,0183	0,6942

Gambar 4 menampilkan perbandingan F1-*weighted* dan distribusi nilai per *fold* dari keempat skenario, memperlihatkan pola konsistensi dan variabilitas masing-masing pendekatan secara lebih rinci.



Gambar 4. *Perbandingan F1-weighted dan distribusi per fold keempat skenario*

Skenario tanpa penyeimbangan tambahan menghasilkan *F1-weighted* tertinggi (0,6951 *sensitive learning* tidak perlu mengkompensasi bias distribusi, dan *focal loss* kehilangan keunggulannya karena tidak ada kelas yang secara konsisten "sulit" akibat

ketimpangan jumlah. Variasi standar deviasi terbesar justru ada pada *focal loss* (0,0183), mengindikasikan perilaku yang sedikit lebih tidak stabil antar *fold*; ini merupakan pertimbangan praktis ketika stabilitas prediksi lebih diprioritaskan daripada performa puncak.), namun selisih terhadap ketiga skenario lainnya tidak melebihi 0,003 poin. Ini bukan hasil yang mengejutkan jika dilihat dari desain eksperimen: sampel yang sudah diseimbangkan secara stratifikasi mengeliminasi sumber imbalance utama sebelum teknik balancing tambahan sempat bekerja. Dengan proporsi kelas yang setara 1:1:1, random oversampling tidak menambah informasi baru, *cost- sensitive learning* tidak perlu mengkompensasi bias distribusi, dan *focal loss* kehilangan keunggulannya karena tidak ada kelas yang secara konsisten "sulit" akibat ketimpangan jumlah. Variasi standar deviasi terbesar justru ada pada *focal loss* (0,0183), mengindikasikan perilaku yang sedikit lebih tidak stabil antar *fold*; ini merupakan pertimbangan praktis ketika stabilitas prediksi lebih diprioritaskan daripada performa puncak.

Uji Signifikansi Statistik (*Wilcoxon*)

Untuk memastikan perbedaan antar metode bukan sekadar fluktuasi sampling, dilakukan uji *Wilcoxon signed-rank* pada nilai *F1-weighted 10 fold* dari semua enam pasangan skenario (α=0,05).

Tabel 3. Hasil Uji *Wilcoxon Signed-Rank* Antarpassangan Skenario

Pasangan	p-value	*Effect size (r)*	Kesimpulan
Tanpa *balancing* vs *Oversampling*	0,4922	0,217 (kecil)	Tidak signifikan
Tanpa *balancing* vs *Cost-sensitive*	0,5566	0,186 (kecil)	Tidak signifikan
Tanpa *balancing* vs *Focal loss*	0,4922	0,217 (kecil)	Tidak signifikan
Oversampling vs *Cost-sensitive*	0,9219	0,031 (sangat kecil)	Tidak signifikan
Oversampling vs *Focal loss*	1,0000	0,000	Tidak signifikan
Cost-sensitive vs *Focal loss*	0,8457	0,062 (sangat kecil)	Tidak signifikan

Tabel 3 menyajikan hasil uji *Wilcoxon Signed-Rank* untuk setiap pasangan skenario beserta nilai *effect size* dan kesimpulan statistiknya. Tidak ada satu pasangan pun yang

menghasilkan p < 0,05. Pasangan *oversampling* vs *focal loss* mencatat p=1,000 dengan *effect size* nol. Hal ini menujukkan kedua skenario menghasilkan profil F1 per *fold* yang identik secara praktis. *Effect size* terbesar hanya 0,217 (kategori kecil menurut skala Cohen), memperkuat simpulan bahwa perbedaan pada Tabel 2 tidak memiliki signifikansi statistik maupun praktis. Pada sampel yang sudah seimbang secara desain, pemilihan teknik balancing tidak menentukan kualitas model secara keseluruhan. Namun demikian, sebagaimana ditunjukkan berikutnya, efeknya tetap terlihat di level kelas minoritas.

Evaluasi per Kelas (*Classification Report*)

Meski metrik agregat hampir identik, analisis per kelas mengungkap nuansa yang relevan secara praktis.

Tabel 4. *Classification Report* per Kelas pada Model Representatif Tiap Skenario

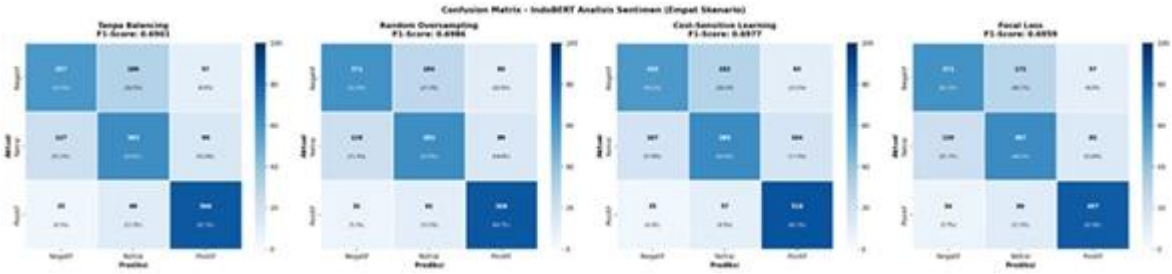
Skenario	Kelas	*Precision*	*Recall*	F1-*score*
Tanpa *balancing*	Negatif	0,7216	0,6133	0,6631
	Netral	0,5943	0,63	0,6117
	Positif	0,7737	0,8433	0,807
Oversampling	Negatif	0,7262	0,6367	0,6785
	Netral	0,6164	0,6533	0,6343
	Positif	0,7806	0,83	0,8045
Cost-sensitive	Negatif	0,6837	0,67	0,6768
	Netral	0,6303	0,5967	0,613
	Positif	0,7764	0,8333	0,8039
Focal loss	Negatif	0,7245	0,64	0,6796
	Netral	0,6226	0,6433	0,6328
	Positif	0,7692	0,8333	0,8

Tabel 4 menyajikan classification report per kelas pada model representatif dari tiap skenario, mencakup nilai *precision, recall,* dan F1-*score* untuk masing-masing kelas sentimen. Kelas positif konsisten mendapatkan F1-*score* tertinggi di semua skenario, yaitu berkisar antara 0,80 hingga 0,81, mencerminkan representasi linguistik yang lebih seragam pada ulasan apresiasi. Kelas netral adalah yang paling sulit dikenali model: F1-*score* terendah 0,61 hingga 0,63 dengan precision lemah, terutama pada skenario tanpa *balancing*

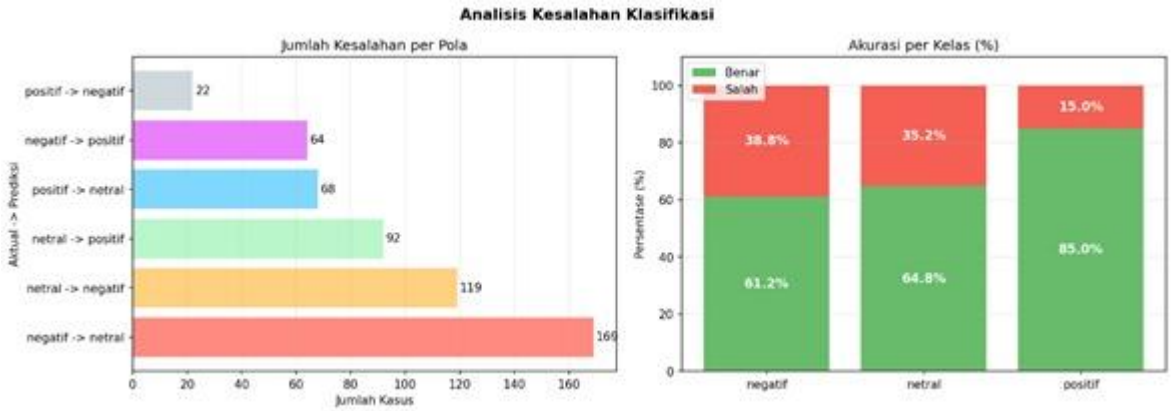
(0,5943), mencerminkan ambiguitas semantik ulasan singkat di perbatasan sentimen. Yang signifikan untuk dicatat adalah teknik balancing memperlihatkan peningkatan kecil namun konsisten pada kelas minoritas; *oversampling* menaikkan F1-*score* kelas netral dari 0,6117 ke 0,6343, sementara *cost-sensitive* meningkatkan *recall* kelas negatif dari 0,6133 ke 0,6700. Peningkatan ini tidak cukup besar untuk menggeser metrik agregat secara signifikan, namun bermakna dalam konteks aplikasi di mana identifikasi keluhan (kelas negatif) memiliki nilai actionable lebih tinggi daripada sekadar mengkonfirmasi ulasan positif.

Analisis Kesalahan Klasifikasi

Pada model tanpa balancing dengan 1.800 sampel uji, 70,3% prediksi benar dan 29,7% salah. Pola kesalahan tidak acak; ada struktur yang mencerminkan tantangan linguistik spesifik. Gambar 5 menampilkan *confusion matrix* model tanpa *balancing* pada *test set* (n=1.800), yang memvisualisasikan sebaran prediksi benar dan salah untuk setiap kombinasi kelas secara menyeluruh. Gambar 6 memperlihatkan distribusi jenis kesalahan klasifikasi berdasarkan pasangan kelas, sehingga pola kesalahan yang dominan dapat diidentifikasi secara lebih intuitif.



***Gambar 5**. Confusion matrix model tanpa balancing pada test set (n=1.800)*



***Gambar 6.** Distribusi jenis kesalahan klasifikasi berdasarkan pasangan kelas*

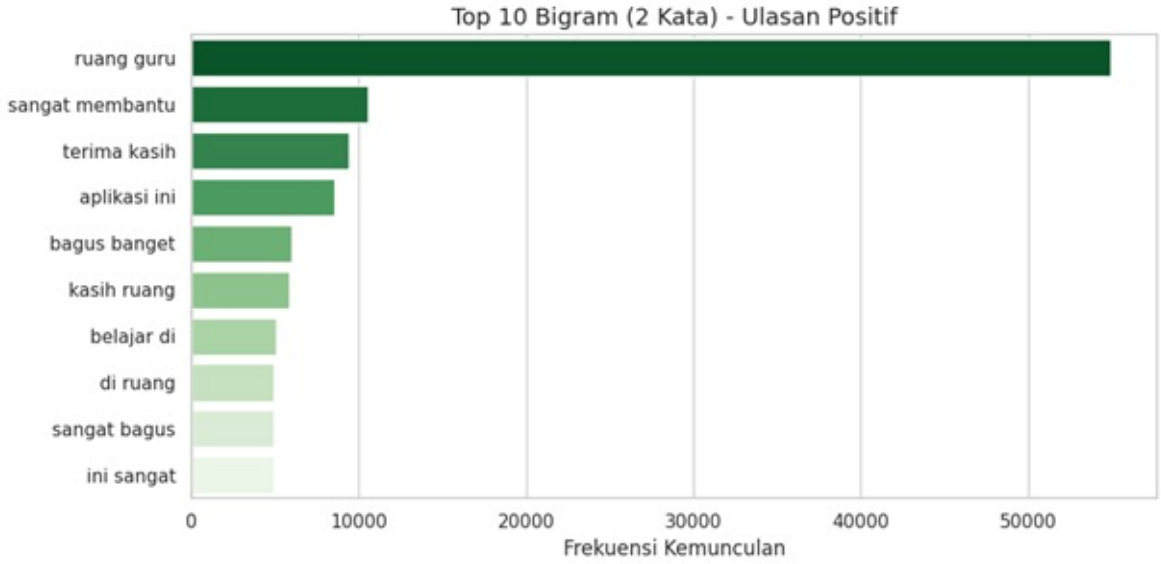
Kesalahan terbesar adalah negatif yang diprediksi netral (169 kasus) dan netral yang diprediksi negatif (119 kasus); keduanya menunjukkan bahwa batas semantik antara ekspresi tidak puas dan ekspresi ambigu masih kabur bagi model. Ulasan singkat seperti "bagus" atau

"baguss" yang seharusnya netral kerap diprediksi positif karena model mengandalkan sinyal leksikal permukaan. Lebih kritis lagi, ulasan campuran yang memuji satu fitur sekaligus mengeluhkan kuota atau bug, yang secara keseluruhan seharusnya negatif atau netral, sering salah klasifikasi karena model belum cukup menangkap negasi dan kontras dalam satu kalimat majemuk. Sarkasme dan kritik yang dibungkus diksi positif masih terprediksi positif; ini merupakan sebuah kelemahan yang umum bahkan pada model berbasis BERT tanpa data pelatihan sarkasme yang memadai. Distribusi kesalahan ini mengonfirmasi bahwa hambatan utama bukan lagi imbalance kelas, yang sudah diatasi di level desain, melainkan kompleksitas linguistik ulasan pengguna berbahasa Indonesia yang informal, kaya kontras, dan kerap ironis.

Implikasi terhadap Kualitas Layanan Pendidikan Digital

Analisis sentimen ini berfungsi sebagai instrumen evaluasi yang memberikan lapisan konteks kualitatif melalui *wordcloud* dan *bigram,* dimana temuan ini tidak tertangkap oleh angka F1-*score* semata.

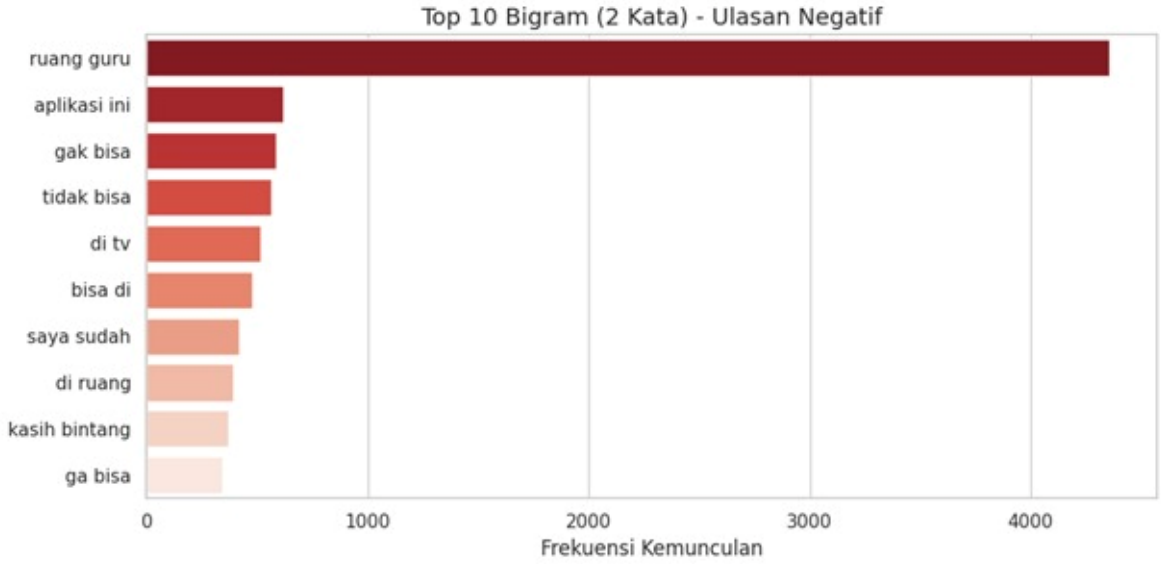
Gambar 7 menampilkan sepuluh *bigram* paling frekuen pada ulasan positif, di mana pasangan kata "sangat membantu" (lebih dari 10.000 kemunculan) dan "terima kasih" mendominasi daftar ini, mengindikasikan bahwa apresiasi pengguna bersifat generik tanpa menyebut fitur spesifik. Kondisi ini justru membatasi nilai diagnostiknya bagi tim pengembang karena tidak memberikan informasi yyang cukup tentang aspek mana dari aplikasi yang paling diunggulkan pengguna.



Gambar 7. *Sepuluh bigram paling frekuen pada ulasan positif Ruang Guru.*

Gambar 8 memperlihatkan bahwa bigram dominan pada ulasan negatif adalah "gak bisa", "tidak bisa", dan "di tv". Kemunculan "gak/tidak/ga bisa" dengan sekitar 1.500 kemunculan gabungan merupakan sinyal kuat bahwa fungsionalitas yang tidak berjalan adalah keluhan paling dominan. Sementara token "di tv" mengindikasikan masalah spesifik pada fitur screen-mirroring atau pemutaran konten melalui televisi, yang merupakan use case yang tampaknya belum dioptimalkan oleh pengembang.

Pola kesalahan model memperkuat gambaran ini, di mana kebingungan antara kelas negatif dan netral kemungkinan bersumber dari gaya komunikasi tidak langsung yang umum dalam konteks budaya Indonesia. Artinya, bahkan ulasan yang salah diklasifikasikan oleh model tetap membawa sinyal yang perlu diperhatikan karena ulasan tersebut kemungkinan mengandung sentimen lemah yang membutuhkan perhatian manual lebih lanjut. Secara operasional, sistem ini paling efektif jika difokuskan pada ulasan yang diprediksi negatif dengan *confidence* tinggi karena *precision* model pada kelas ini mencapai 0,72 sehingga layak dijadikan prioritas tindak lanjut, sementara prediksi di zona abu-abu negatif/netral dengan *confidence* rendah lebih tepat dieskalasi untuk *review* manual. Integrasi *pipeline* ini memungkinkan tim produk Ruangguru mengidentifikasi isu berulang seperti *bug streaming*, keluhan kuota, dan masalah langganan secara sistematis dari ratusan ribu ulasan tanpa pembacaan manual penuh.



Gambar 8. Sepuluh bigram paling frekuen pada ulasan negatif Ruang Guru.

KESIMPULAN

Penelitian ini berhasil mengembangkan sistem analisis sentimen berbasis IndoBERT-base-p1 untuk mengklasifikasikan ulasan pengguna Ruangguru ke dalam tiga

kelas sentimen menggunakan kerangka 10-Fold Stratified Cross-Validation. Dari empat skenario penanganan *class imbalance* yang diujicobakan, skenario tanpa penyeimbangan tambahan menghasilkan F1-*weighted* terbaik sebesar 0,6951, namun selisih terhadap ketiga skenario lainnya tidak melebihi 0,003 poin dan tidak signifikan secara statistik berdasarkan uji Wilcoxon Signed-Rank ($p > 0,05$). Temuan ini menunjukkan bahwa pada dataset yang sudah diseimbangkan secara stratifikasi 1:1:1, teknik *random oversampling*, *cost-sensitive learning*, dan *focal loss* tidak memberikan peningkatan yang bermakna pada metrik agregat.

Analisis per kelas mengungkap bahwa hambatan utama model bukan lagi ketidakseimbangan distribusi data, melainkan kompleksitas linguistik ulasan berbahasa Indonesia yang informal dan kerap ironis. Kelas netral menjadi kelas paling sulit dikenali dengan F1-score berkisar 0,61 hingga 0,63, dan kesalahan terbesar terjadi pada batas semantik antara kelas negatif dan netral yang dipicu oleh penggunaan bahasa tidak langsung dan sarkasme yang umum dalam konteks budaya Indonesia. Untuk penelitian selanjutnya, disarankan eksplorasi augmentasi data berbasis parafrase, pemanfaatan IndoBERT-large, serta penambahan data berlabel sarkasme dan ulasan campuran untuk meningkatkan kemampuan model dalam menangani nuansa linguistik bahasa Indonesia yang lebih kompleks.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Sistem Informasi Universitas Darwan Ali atas dukungan akademik yang diberikan dalam pelaksanaan penelitian ini. Apresiasi juga disampaikan peneydia datset publik ulasan aplikasi Ruangguru di platform Kaggle yang telah memungkinkan penelitian ini terlaksana. Penulis juga berterima kasih kepada seluruh pihak yang telah memberikan masukan dan dukungan dalam proses penyelesaian penelitian ini.

DAFTAR PUSTAKA

Antariksa, M. D. S., Sugiharto, A., & Surarso, B. (2025). Fine-Tuning Bert for Scientific Document Classification in Computer and Informatics Fields. *2025 12th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 1–6. https://doi.org/10.1109/ICITACEE66165.2025.11232930

Cao Lu and Liu, X. and S. H. (2022). Adaptable Focal Loss for Imbalanced Text Classification. In Y. and Z. Y. and X. N. and A. H. R. and F. G. and G. A. and M. M.

Shen Hong and Sang (Ed.), *Parallel and Distributed Computing, Applications and Technologies* (pp. 466–475). Springer International Publishing.

Dwi Purnomo, T., & Sutopo, J. (2024). *COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA*. (3), 11–21. https://doi.org/10.56127/ijst.v3i3.1

Jaiswal, A., & Milios, E. (2023). *Breaking the Token Barrier: Chunking and Convolution for Efficient Long Text Classification with BERT*. http://arxiv.org/abs/2310.20558

Kaope, C., & Pristyanto, Y. (2023). The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 227–238. https://doi.org/10.30812/matrik.v22i2.2515

Kyriakidis, A., & Tsafarakis, S. (2025). Extracting knowledge from customer reviews: an integrated framework for digital platform analytics. *International Transactions in Operational Research*, *32*(4), 2061–2086. https://doi.org/10.1111/itor.13537

Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., Bonakdarpour, B., Vassar, R., Shen, L., Jiang, G., Wang, F., Pathak, J., & Luo, Y. (2023). AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *Journal of Biomedical Informatics*, *144*. https://doi.org/10.1016/j.jbi.2023.104442

Nasution, A. K. P. (2022). Education Technology Research Trends in Indonesia during the COVID-19 Pandemic. *Asia Pacific Journal of Educators and Education*, *36*(2), 65–76. https://doi.org/10.21315/apjee2021.36.2.4

Prasetiadi, A., Dwi Sripamuji, A., Riski Amalia, R., Saputra, J., & Ramadhanti, I. (2024). Automatic Vocal Completion for Indonesian Language Based on Recurrent Neural Network. *IT Journal Research and Development*, *9*(1), 14–26. https://doi.org/10.25299/itjrd.2024.14171

Setyo Nugroho, K., Yullian Sukmadewa, A., Wuswilahaken, H. D., Abdurrachman Bachtiar, F., & Yudistira, N. (n.d.). *BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews*. Retrieved https://research.google/teams/brain.

Tapio, R. P. (2025). The Role of Data Assumptions in Selecting Between Parametric and Nonparametric Tests. *Asian Journal of Probability and Statistics*, *27*(11), 127–135. https://doi.org/10.9734/ajpas/2025/v27i11830

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-05791-7