ANALISIS SENTIMEN DRIVER TERHADAP KINERJA APLIKASI DAN KEBIJAKAN LALAMOVE: STUDI KOMPARASI MODEL COMPLEMENT NAÏVE BAYES DAN INDOBERT

David Marlon Hermando[1*], Ika Nur Laily Fitriana[2]
[1,2]*Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas Terbuka, Tangerang Selatan, Banten, Indonesia*

**Penulis Korespondensi: davidhermando7@gmail.com*

ABSTRAK

Ulasan mitra pengemudi pada platform Google Play Store merupakan sumber data yang kaya namun belum dimanfaatkan secara optimal untuk mengevaluasi kualitas layanan aplikasi logistik berbasis *sharing economy* seperti Lalamove. Penelitian ini bertujuan untuk melakukan analisis sentimen dan ekstraksi opini terhadap ulasan mitra pengemudi Lalamove guna membandingkan kinerja algoritma Complement Naïve Bayes dan model *deep learning* IndoBERT dalam klasifikasi sentimen berbahasa Indonesia. Data dikumpulkan melalui teknik *web scraping* dari Google Play Store sebanyak 10.000 ulasan, kemudian dilabeli secara otomatis berdasarkan skor bintang. Selanjutnya, sebanyak 500 sampel ulasan divalidasi secara manual untuk mengoreksi anomali polaritas antara *rating* dan narasi teks. Prapemrosesan meliputi *case folding*, *cleansing*, normalisasi leksikal, dan *stopword removal* khusus untuk jalur Naïve Bayes. Data dibagi dengan rasio 80:20 untuk pelatihan dan pengujian. Naïve Bayes dikombinasikan dengan ekstraksi fitur TF-IDF dan penyeimbangan kelas menggunakan SMOTE, sedangkan IndoBERT dioptimasi melalui *fine-tuning* dengan pembobotan kelas pada fungsi kerugian. Dari 8.738 data valid, distribusi sentimen didominasi oleh kelas Negatif (68,4%), diikuti Positif (28,2%), dan Netral (3,4%), yang mengonfirmasi adanya ketidakseimbangan kelas yang nyata. IndoBERT mencapai akurasi 87,47% dan *Weighted* F1 87,60%, mengungguli Complement Naïve Bayes dengan akurasi 82,44% dan *Weighted* F1 84,77%. Penelitian ini menyimpulkan bahwa IndoBERT merupakan model yang lebih unggul untuk analisis sentimen ulasan mitra pengemudi berbahasa Indonesia, dengan kemampuan memahami konteks kalimat informal dan *slang* yang jauh lebih baik dibandingkan pendekatan berbasis frekuensi kata. Integrasi IndoBERT direkomendasikan sebagai fondasi sistem pemantauan opini (*opinion mining*) berskala besar pada platform logistik digital di Indonesia.

Kata kunci: analisis sentimen, IndoBERT, Naïve Bayes, *text mining*, ulasan pengemudi

1. PENDAHULUAN

Pertumbuhan sektor logistik *on-demand* di Indonesia bertumpu pada mitra pengemudi sebagai pelaku utama operasional. Namun, inovasi fitur dan kebijakan platform tidak selalu selaras dengan ekspektasi mitra, sehingga keluhan terhadap kebijakan perusahaan banyak disuarakan pada kanal publik seperti kolom ulasan toko aplikasi (Destika et al., 2026). Evaluasi sentimen mitra pengemudi karenanya diperlukan untuk memperbaiki kualitas layanan dan menjaga keberlanjutan kemitraan.

Dalam konteks layanan aplikasi Lalamove, ribuan ulasan pengguna di Google Play Store merupakan sumber data yang relevan untuk dianalisis. Penelitian terdahulu oleh Riyana (2026) melakukan penambangan data (*text mining*) terhadap ulasan Lalamove menggunakan berbagai algoritma *Machine Learning* tradisional, seperti Naïve Bayes, *Support Vector Machine* (SVM), dan *Random Forest*. Meskipun algoritma seperti SVM mampu memberikan hasil klasifikasi yang baik, penelitian tersebut secara eksplisit merekomendasikan penggunaan model berbasis

transformer mutakhir seperti IndoBERT untuk penelitian di masa mendatang guna mendapatkan pemahaman konteks semantik yang lebih akurat (Riyana, 2026).

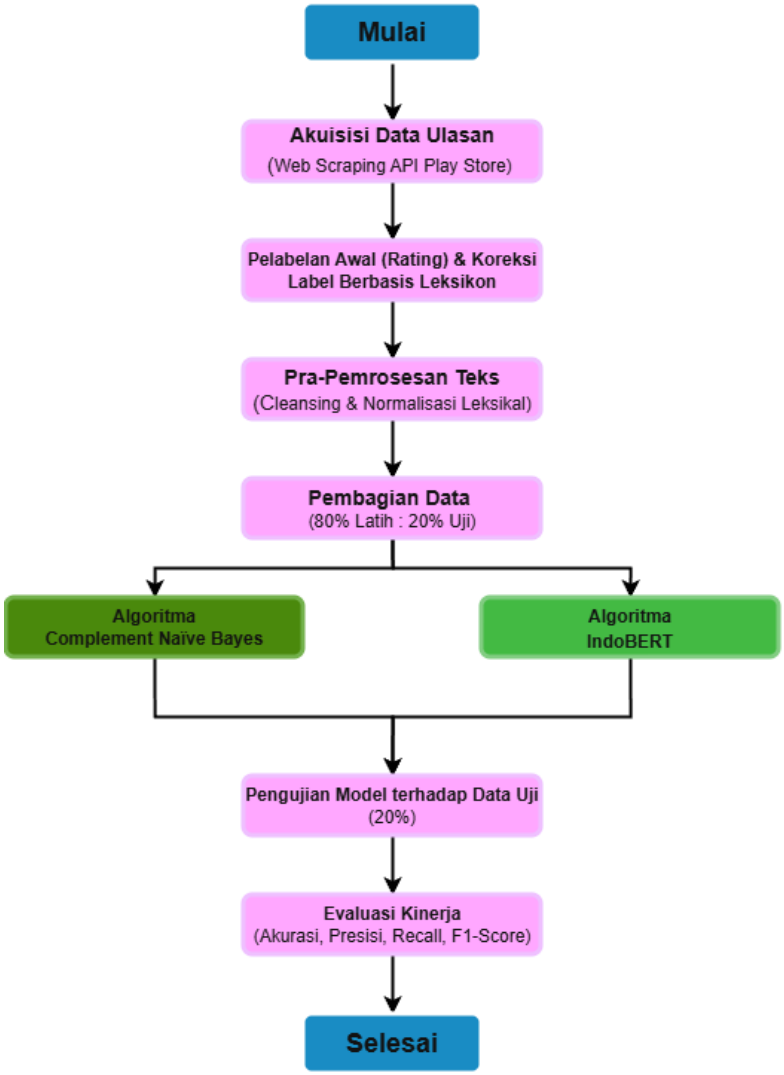
Tantangan lain dalam analisis sentimen ulasan adalah ketidakseimbangan kelas (*imbalanced data*), ketika sentimen negatif mendominasi korpus. Naïve Bayes standar cenderung bias terhadap kelas mayoritas pada kondisi tersebut. Varian Complement Naïve Bayes (CNB) dirancang untuk mengurangi bias ini dengan menghitung probabilitas komplemen setiap kelas (Rennie et al., 2003), dan penerapannya pada klasifikasi teks berbahasa Indonesia dilaporkan memberikan hasil yang kompetitif (Randhika et al., 2021). Pada data tidak seimbang, CNB juga dilaporkan menghasilkan akurasi dan presisi yang lebih stabil dibandingkan varian Multinomial, Gaussian, maupun Bernoulli (Sani et al., 2022).

Di sisi lain, pendekatan *deep learning* menawarkan penanganan variasi linguistik yang lebih baik. IndoBERT dilaporkan efektif menganalisis opini pengguna aplikasi pada Google Play Store dengan akurasi klasifikasi mencapai 90% (Andhika et al., 2025). Kemampuan tersebut bersumber dari mekanisme *self-attention* yang memodelkan konteks kata dalam keseluruhan kalimat, serta tokenisasi WordPiece yang memungkinkan penanganan kata tidak baku, kesalahan ejaan, dan ragam gaul (*out-of-vocabulary*) (Andhika et al., 2025).

Berdasarkan rekomendasi dan celah dari penelitian terdahulu, studi ini bertujuan untuk mengekstraksi sentimen mitra pengemudi terhadap kinerja dan kebijakan aplikasi Lalamove. Penelitian ini akan membandingkan secara komprehensif performa algoritma Complement Naïve Bayes sebagai *baseline* probabilistik yang teruji pada data timpang, dengan model *deep learning* IndoBERT sebagai *state-of-the-art* dalam pemrosesan bahasa alami berbahasa Indonesia.

2. METODE

Penelitian ini menerapkan desain studi komparatif kuantitatif untuk mengevaluasi dan membandingkan performa berbagai algoritma dalam klasifikasi sentimen teks. Untuk memberikan gambaran yang komprehensif dan sistematis mengenai alur kerja penelitian, tahapan-tahapan yang dilakukan mulai dari pengumpulan data, prapemrosesan, pemodelan, hingga evaluasi disajikan pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

2.1. Akuisisi Data dan Pelabelan

Korpus data yang digunakan merupakan data sekunder berupa ulasan pengguna pada aplikasi Lalamove Driver yang diperoleh melalui teknik web scraping dari platform Google Play Store. Ekstraksi komputasi dijalankan menggunakan pustaka `google-play-scraper` dengan parameter penarikan diatur berdasarkan urutan ulasan terbaru (*sort by newest*) dari Februari 2025 hingga Mei 2026. Sebanyak 10.000 ulasan ditarik sebagai sampel observasi guna menyediakan basis data yang representatif untuk analisis linguistik dan pemodelan komputasi.

Proses pelabelan awal (*ground truth*) dilakukan secara otomatis berdasarkan metrik skor bintang (rating) yang diberikan oleh pengguna. Parameter algoritmik menetapkan bahwa ulasan dengan skor 4 dan 5 diklasifikasikan ke dalam kelas sentimen Positif, skor 3 sebagai Netral, sedangkan skor 1 dan 2 dikategorikan sebagai kelas Negatif.

2.2. Alat dan Perangkat Lunak

Eksperimen komputasi dalam penelitian ini dijalankan secara komprehensif menggunakan bahasa pemrograman Python 3. Seluruh pemrosesan dan pelatihan algoritma dieksekusi melalui lingkungan kerja berbasis *cloud* Google Colaboratory. Guna menangani beban komputasi dari arsitektur *deep learning* yang kompleks, eksekusi pemodelan diakselerasi menggunakan *Graphics Processing Unit* (GPU) tipe T4.

Implementasi teknis pada penelitian ini didukung oleh beberapa pustaka (*library*) khusus yaitu, `google-play-scraper` untuk akuisisi data ulasan secara otomatis, `pandas` dan `numpy` untuk manipulasi dan eksplorasi *data frame*, `Sastrawi` untuk tahapan pra-pemrosesan teks berbahasa Indonesia, khususnya penghapusan *stopword*, `scikit-learn` untuk ekstraksi fitur TF-IDF, pembagian dataset, metrik evaluasi, serta pemodelan algoritma Complement Naïve Bayes, `imbalanced-learn` untuk implementasi teknik penyeimbangan kelas melalui algoritma SMOTE, serta `transformers` dan `torch` dari HuggingFace untuk memfasilitasi tokenisasi dan proses *fine-tuning* pada arsitektur model IndoBERT.

2.3. Validasi Manual dan Koreksi Pelabelan (*Lexicon-Based Auto-Correction*)

Metrik skor bintang tidak selalu selaras dengan narasi teks ulasan, sehingga dilakukan pengujian validitas pelabelan awal. Sebanyak 500 ulasan diambil secara acak berstrata (*stratified sampling*) untuk diperiksa secara manual. Pemeriksaan ini digunakan untuk mengukur tingkat ketidaksesuaian antara skor bintang dan polaritas teks; hasilnya dilaporkan pada bagian Hasil dan Pembahasan. Kecenderungan keluhan mitra pengemudi terhadap kebijakan platform juga dilaporkan pada studi terdahulu (Destika et al., 2026).

Tingkat kesalahan pelabelan awal tersebut memerlukan koreksi sebelum pemodelan. Pelabelan otomatis berbasis leksikon (*lexicon-based*) dilaporkan efektif untuk membangun *ground truth* pada data ulasan aplikasi sekaligus mengurangi beban pelabelan manual (Asri et al., 2022).

Mengadaptasi efektivitas pendekatan tersebut, penelitian ini merancang sebuah mekanisme koreksi kebaruan berbasis leksikon (*Lexicon-Based Auto-Correction*) spesifik yang diimplementasikan pada keseluruhan korpus data. Proses koreksi ini mengacu pada aturan leksikal ketat (*strict rule-based*): Jika sebuah ulasan berlabel awal Positif atau Netral namun terdeteksi mengandung sekurang-kurangnya satu kata kunci dari kamus keluhan operasional (seperti ‘potongan’, ‘sepi’, ‘argo’, ‘suspend’, atau ‘fiktif’), maka label tersebut secara paksa diubah (*override*) menjadi Negatif. Pendekatan *hybrid* ini memastikan validitas sentimen secara semantik dan mencegah masuknya data cacat (*garbage in*) ke dalam tahapan pemodelan *Machine Learning*.

2.4. Pra-Pemrosesan Teks

Prapemrosesan dilakukan untuk mengubah ulasan tekstual tidak terstruktur menjadi teks bersih (*clean text*) sebelum direpresentasikan secara numerik (Riyana, 2026). Prosedur dasar meliputi *case folding*, *cleaning* untuk menghapus tautan, angka, simbol, dan tanda baca, *tokenizing*, serta *stemming* (Riyana, 2026). Tahap normalisasi turut diterapkan untuk memperbaiki kesalahan eja dan mengembalikan singkatan ke bentuk bakunya (Kusnaya et al., 2025). Setelah pembersihan leksikal dasar, data dipilah menjadi dua jalur prapemrosesan sesuai arsitektur pemodelan.

2.5. Ekstraksi Fitur dan Pemodelan

Himpunan data yang telah melalui tahapan prapemrosesan dan normalisasi leksikal selanjutnya dipartisi menjadi data latih (*training set*) dan data uji (*testing set*) dengan rasio proporsional 80:20. Pembagian ini dieksekusi secara berstrata (*stratified split*) guna menjamin bahwa distribusi label sentimen (Negatif, Netral, dan Positif) tetap konsisten dan seimbang baik pada partisi data latih maupun data uji. Setelah pembagian data selesai, pipa pemrosesan dipisahkan secara biner ke dalam dua skenario arsitektur pemodelan yang berbeda.

2.5.1. Algoritma Complement Naïve Bayes

Pada jalur pemodelan pertama, data teks terlebih dahulu melalui tahapan eliminasi kata hubung (*stopword removal*) menggunakan bantuan pustaka Sastrawi. Guna menjaga integritas

makna dari ulasan keluhan operasional mitra pengemudi, aturan pengecualian diterapkan secara ketat terhadap kata-kata negasi dan indikator sentimen krusial (meliputi: 'tidak', 'belum', 'bukan', 'enggak', 'jangan', 'tanpa', 'kurang', 'jarang', 'hampir', 'nyaris', 'susah', 'sulit', 'gagal', 'buruk', dan 'jelek') agar tidak ikut terhapus.

Teks yang telah difilter kemudian ditransformasikan ke dalam representasi numerik menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Representasi numerik vektor dibentuk melalui konfigurasi ekstraksi fitur dengan menangkap kombinasi kata tunggal dan ganda (`ngram_range=(1, 2)`). Pembobotan TF-IDF untuk setiap kata ke-i pada dokumen ke-j dihitung menggunakan persamaan (Kusnaya et al., 2025):

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

(Keterangan: $W_{i,j}$ = Bobot kata ke-i pada dokumen ke-j, $tf_{i,j}$ = Jumlah kemunculan kata ke-i pada dokumen ke-j, N = Total seluruh ulasan, df_i = Jumlah dokumen yang mengandung kata ke-i)

Ambang frekuensi dokumen minimum ditetapkan sehingga sebuah token harus muncul pada sekurang-kurangnya dua dokumen (min_df=2), dengan batas maksimum 95% (max_df=0.95) untuk mengabaikan token yang terlalu umum. Penskalaan frekuensi logaritmik (sublinear_tf=True) diterapkan untuk meredam bias pada istilah yang berulang dalam satu dokumen. Ekstraksi kosakata (*fit*) dilakukan hanya pada data latih untuk mencegah kebocoran informasi (*data leakage*).

Untuk menangani ketidakseimbangan kelas (*class imbalance*), teknik *Synthetic Minority Over-sampling Technique* (SMOTE) diterapkan pada ruang vektor fitur TF-IDF (Chawla et al., 2002). SMOTE mensintesis sampel minoritas baru melalui interpolasi linear antara suatu sampel minoritas dan tetangganya berdasarkan prinsip *k-Nearest Neighbors* (k-NN), dengan jumlah tetangga disesuaikan terhadap ukuran kelas minoritas yang tersedia.

Data latih yang telah seimbang tersebut kemudian diterapkan pada algoritma Complement Naïve Bayes (CNB). Klasifikasi dokumen (d) ditentukan berdasarkan probabilitas komplemen, yaitu probabilitas sebuah kata yang kemunculannya diketahui berada pada kelas bukan c atau komplemennya (c') melalui formulasi fungsi objektif (Rennie et al., 2003; Randhika et al., 2021):

$$CNB(d) = \arg\max_{c \in C}\left[\log P(c) + \sum\nolimits_{1 \le k \le n_d} \log P(x_k|c')\right] \quad (2)$$

Melalui pendekatan adaptasi dataset tidak seimbang ini, perhitungan bobot komplemen diselesaikan dengan melakukan estimasi parameter probabilitas $\hat{\theta}_{ci}$ dan penugasan bobot fitur w_{ci} melalui persamaan matematis:

$$\hat{\theta}_{ci} = \frac{\alpha + \sum_{j:y_j \neq c} d_{ij}}{\alpha|V| + \sum_{j:y_j \neq c}\sum_k d_{kj}} \quad (3)$$

$$w_{ci} = \log \hat{\theta}_{ci} \quad (4)$$

Di mana d_{ij} merepresentasikan nilai bobot TF-IDF untuk fitur i dalam dokumen j yang berada di luar kelas c, sedangkan α bertindak sebagai parameter pemulusan (*smoothing parameter*). Pendekatan ini memastikan klasifikasi berjalan tanpa bias ke arah kelas mayoritas (Sani et al., 2022).

2.5.2. Arsitektur *Deep Learning* IndoBERT

Klasifikasi kedua menggunakan model IndoBERT (`indobenchmark/indobert-base-p1`) melalui proses *fine-tuning* (Wilie et al., 2020). Model ini dirancang untuk memproses token masukan guna menangkap konteks dan makna ulasan secara dua arah, sehingga memberikan pemahaman semantik yang jauh lebih mendalam (Salsabila et al., 2025).

Proses tokenisasi teks dieksekusi menggunakan teknik *WordPiece* bawaan model IndoBERT dengan pembatasan panjang sekuens maksimal sebesar 128 *token* untuk mengefisiensikan konsumsi VRAM perangkat keras. Model bahasa yang digunakan ini secara arsitektural memiliki komponen *12 hidden layers* yang telah melalui pelatihan awal (*pre-trained*) di atas 220 juta kata berbahasa Indonesia, sehingga mampu mengenali variasi semantik informal (Andhika et al., 2025). Kekuatan utama IndoBERT bersandar pada mekanisme *self-attention* dua arah (*bidirectional*) yang memungkinkan pemahaman konteks kata secara simultan dari keseluruhan kalimat, dengan formulasi matematika dihitung sebagai berikut (Vaswani et al., 2017; Andhika et al., 2025):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Di mana matriks Q (*Query*), K (*Key*), dan V (*Value*) merupakan bentuk representasi ruang vektor yang dibentuk secara otomatis dari sekuens *token* teks masukan, dan d_k melambangkan dimensi dari vektor *key*.

Berbeda dengan *oversampling* pada ruang fitur TF-IDF, ketidakseimbangan kelas pada pipa IndoBERT ditangani pada level algoritma melalui pembobotan kelas (*Class Weights*). Kedua pendekatan memiliki tujuan yang sama, yaitu mengurangi bias terhadap kelas mayoritas, tetapi berbeda pada mekanismenya: SMOTE menambah sampel minoritas pada tingkat data sebelum pelatihan, sedangkan *Class Weights* menyesuaikan fungsi kerugian (*loss function*) dengan penalti lebih besar pada kesalahan kelas minoritas tanpa mengubah distribusi data (Setiawan et al., 2025). *Class Weights* dipilih karena kompatibel dengan arsitektur *Transformer* dan tidak memerlukan augmentasi tingkat teks yang berpotensi mengganggu urutan token (Madabushi et al., 2019).

Untuk mengatasi ketidakseimbangan representasi kelas sentimen pada data latih, bobot kelas dihitung secara inversi proporsional terhadap frekuensi kemunculannya melalui formulasi berikut:

$$W = \frac{N}{C \times n_c} \quad (6)$$

di mana W_c merupakan bobot untuk kelas c, N adalah total sampel, C adalah jumlah kelas, dan n_c merupakan jumlah sampel pada kelas c. Nilai bobot ini disematkan secara langsung ke dalam fungsi kerugian *Cross-Entropy Loss* saat melakukan iterasi *fine-tuning*. Pendekatan *cost-sensitive learning* ini menetapkan penalti yang lebih besar terhadap kesalahan prediksi pada kelas minoritas (Netral dan Positif), sehingga memaksa jaringan saraf tiruan untuk mengoptimasi pembaruan bobot parameter secara adil selama fase *gradient descent* dan meminimalisasi risiko degradasi performa model (Johnson & Khoshgoftaar, 2019).

Selanjutnya, proses *fine-tuning* dijalankan menggunakan optimasi *AdamW* dengan *learning rate* sebesar 2×10^{-5}, nilai parameter *weight decay* 0,01 untuk mencegah *overfitting*, dan mekanisme *warmup steps* untuk menjaga stabilitas pembaruan bobot pada fase awal pelatihan (Devlin et al., 2019). Sebagai pelengkap untuk memastikan model mencapai performa generalisasi terbaik, metode *Early Stopping* turut diimplementasikan. Metode ini akan menghentikan proses pelatihan secara otomatis apabila tidak terjadi penurunan nilai kerugian (*validation loss*) pada data uji selama dua *epoch* berturut-turut.

2.6. Metode Evaluasi Model

Penilaian terhadap kualitas model klasifikasi diukur menggunakan instrumen *confusion matrix*. Matriks ini berbentuk tabel kontingensi yang memvisualisasikan secara detail perbandingan antara label hasil prediksi algoritma dengan label sebenarnya (*ground truth*) pada data uji (Andhika et al., 2025). Terdapat empat elemen dasar yang menyusun matriks ini: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Keempat

komponen fundamental tersebut menjadi basis perhitungan matematis untuk mengekstrak metrik evaluasi kinerja yang lebih spesifik, yaitu akurasi, presisi, *recall*, dan F1-*score*.

Akurasi merupakan parameter evaluasi standar yang merepresentasikan persentase total tebakan model yang tepat (baik prediksi positif maupun negatif) dibandingkan dengan keseluruhan populasi data uji (Andhika et al., 2025). Formulasi matematis akurasi dihitung melalui persamaan:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Presisi mengukur tingkat ketepatan atau keandalan model ketika memprediksi suatu kelas sentimen. Metrik ini digunakan untuk menilai seberapa banyak data yang ditebak positif oleh algoritma yang secara faktual memang benar data positif, sehingga efektif untuk meminimalisasi deteksi palsu (Andhika et al., 2025). Persamaan presisi didefinisikan sebagai:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall (atau sensitivitas) bertujuan untuk menakar kemampuan algoritma dalam menemukan dan mendeteksi seluruh data aktual yang relevan pada suatu kelas. Dalam konteks ekstraksi ulasan, *recall* menunjukkan rasio ulasan target yang berhasil dikenali oleh sistem (Andhika et al., 2025). Rumus perhitungannya adalah:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

F1-*Score* merupakan titik keseimbangan komputasi yang dihitung melalui rata-rata harmonik antara metrik presisi dan *recall* (Andhika et al., 2025). Metrik ini dinilai lebih representatif dalam mengukur keandalan model dibandingkan akurasi semata, terlebih ketika algoritma sedang dievaluasi pada himpunan data dengan distribusi kelas yang timpang (*imbalanced*). Fungsi matematis F1-*Score* dirumuskan sebagai:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

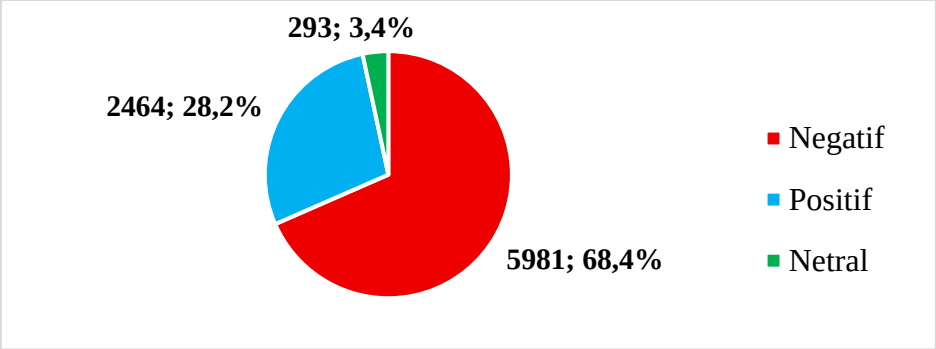
3. HASIL DAN PEMBAHASAN

3.1. Eksplorasi Data

Sebelum dilakukan pemodelan komputasi, tahapan eksplorasi data (*Exploratory Data Analysis*) diterapkan pada ulasan mitra pengemudi Lalamove yang telah diekstraksi. Dari total 10.000 data mentah, proses pembersihan dasar (*cleansing*), pembuangan teks kosong, dan normalisasi leksikal menghasilkan 8.738 baris data valid.

Berdasarkan pengujian validasi silang yang dilakukan secara manual terhadap 500 sampel ulasan, ditemukan adanya ketidaksesuaian polaritas yang cukup besar antara skor bintang (*rating*) dan teks ulasan. Banyak mitra pengemudi memberikan penilaian bintang 5, namun menarasikan keluhan teknis maupun operasional pada teks ulasan (*fake rating*). Secara kuantitatif, hasil validasi menemukan bahwa 14% dari ulasan berlabel Positif dan 50% dari ulasan berlabel Netral secara aktual memuat sentimen Negatif. Sebaliknya, pelabelan pada kelas Negatif konsisten, dengan persentase perbedaan 0%.

Berdasarkan temuan tersebut, koreksi pelabelan berbasis leksikon (*Lexicon-Based Auto-Correction*) sebagaimana diuraikan pada bagian Metode diterapkan pada keseluruhan korpus sebelum pemodelan. Distribusi kelas sentimen akhir menunjukkan ketidakseimbangan data (*imbalanced data*) yang lazim ditemui pada platform ulasan, sebagaimana disajikan pada Gambar 2.



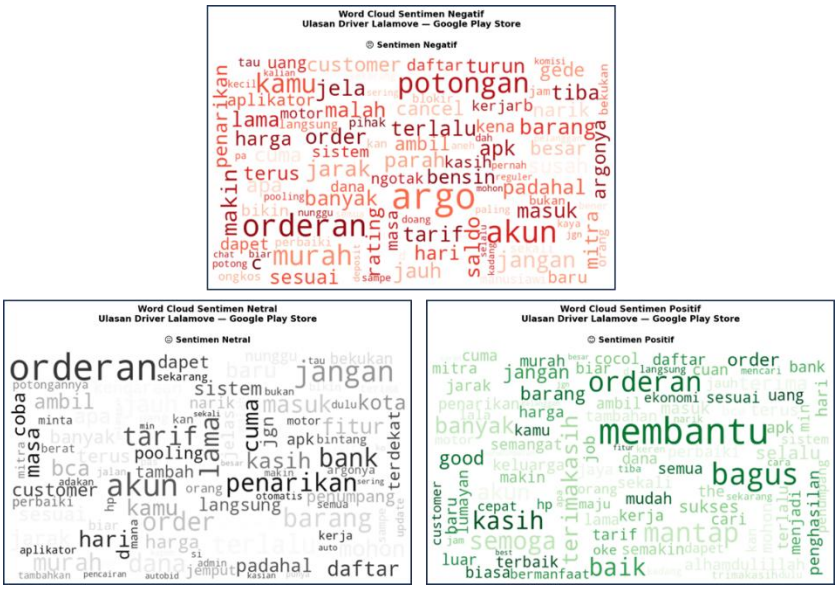
Gambar 2. Distribusi Kelas Sentimen Ulasan Lalamove

Untuk memberikan representasi kualitatif dari masing-masing kelas sentimen, beberapa contoh kutipan teks ulasan asli yang merepresentasikan opini mitra pengemudi disajikan pada Tabel 1.

Tabel 1. Sampel Ulasan Berdasarkan Kelas Sentimen

Label Sentimen	Kutipan Langsung (Setelah *Cleansing*)
Negatif	"harganya tidak ngotak semenjak ada penumpang hrg barang jadi turun siap sepi yang ambil orderannya"
Netral	"untuk no rekening nya mohon di tambahkan selain bank bca agar lebih mudah"
Positif	" keren banget i love lalamove"

Selain itu, guna mendapatkan representasi visual mengenai topik atau kata kunci yang paling dominan, teks ulasan dipetakan ke dalam bentuk *Word Cloud*. Visualisasi komprehensif ini disajikan pada Gambar 3.



Gambar 3. Visualisasi *Word Cloud* Berdasarkan Sentimen

Berdasarkan Gambar 3, dapat diobservasi bahwa pada ulasan bersentimen Negatif, kata-kata yang menonjol umumnya berkaitan dengan keluhan operasional dan finansial seperti "potongan", "sepi", "susah", dan "argo". Pada ulasan Positif, leksikon didominasi oleh apresiasi fungsional aplikasi seperti "membantu", "bagus", dan "mantap". Sementara itu, kelas Netral menampilkan percampuran antara kata yang bersifat fungsional dan pertanyaan teknis. Ekstraksi visual ini mengonfirmasi bahwa mitra pengemudi secara aktif menggunakan ruang

ulasan Play Store untuk merespons kebijakan insentif dan reliabilitas antarmuka aplikasi Lalamove.

Dominasi leksikon finansial dan operasional pada sentimen negatif sejalan dengan kondisi kerja pada ekonomi berbagi (*gig economy*). Mitra pengemudi memiliki posisi tawar (*bargaining power*) yang terbatas dalam penentuan tarif dasar maupun mekanisme distribusi pesanan, sehingga kolom ulasan berfungsi sebagai kanal penyampaian keluhan. Keluhan tersebut berpusat pada transparansi pembagian hasil ('potongan') dan ketidakpastian perolehan pesanan ('sepi'). Temuan ini konsisten dengan studi terdahulu mengenai kontrol platform melalui algoritma, sistem rating, dan potongan pendapatan (Destika et al., 2026).

3.2. Kinerja Model Complement Naïve Bayes

Penutian model Complement Naïve Bayes (CNB) dilakukan menggunakan 20% proporsi data uji, yang merepresentasikan 1.748 ulasan yang tidak dilibatkan dalam proses pelatihan model. Matriks hasil evaluasi komputasi divisualisasikan secara detail melalui *Confusion Matrix* pada Tabel 2.

Tabel 2. *Confusion Matrix* Model *Complement Naïve Bayes*

	Predicted Negative	***Predicted Neutral***	***Predicted Positive***
Actual Negative	1065	97	34
Actual Neutral	33	20	6
Actual Positive	71	66	356

Kinerja agregat dari distribusi prediksi tersebut selanjutnya dikuantifikasi ke dalam metrik evaluasi pada Tabel 3.

Tabel 3. Hasil Evaluasi Model *Complement Naïve Bayes*

Sentimen	***Precision***	***Recall***	***F1-Score***
Negatif	91,10%	89,05%	90,06%
Netral	10,93%	33,90%	16,53%
Positif	89,90%	72,21%	80,09%
Rata-rata bobot	88,06%	82,44%	84,77%

Berdasarkan Tabel 3, CNB dengan TF-IDF dan SMOTE memperoleh akurasi 82,44% pada data uji. Performa antarkelas bervariasi. Pada kelas mayoritas Negatif, model memperoleh *Precision* 91,10% dan *Recall* 89,05% (*F1-Score* 90,06%), yang menunjukkan bahwa fungsi objektif probabilitas komplemen mampu mengenali pola kata keluhan operasional dengan tingkat deteksi palsu (*False Positive*) yang rendah. Pada kelas Positif, model memperoleh *Precision* 89,90%, *Recall* 72,21%, dan *F1-Score* 80,09%. Sebaliknya, performa terendah terjadi pada kelas minoritas Netral, dengan *Precision* 10,93%, *Recall* 33,90%, dan *F1-Score* 16,53%.

Rendahnya presisi pada kelas Netral merupakan konsekuensi dari penyeimbangan kelas. SMOTE meningkatkan *Recall* kelas Netral dari nol, tetapi menurunkan spesifisitas model. Sintesis fitur pada ruang vektor TF-IDF menimbulkan tumpang tindih batas keputusan (*overlapping decision boundaries*), sehingga model cenderung melabeli teks ambigu dari kelas Negatif maupun Positif sebagai Netral dan meningkatkan jumlah *False Positive*. Secara agregat, keterbatasan pada kelas minoritas menurunkan *Macro F1* menjadi 62,23%. Nilai *Weighted F1* tetap 84,77% karena ditopang performa pada kelas dominan.

3.3. Kinerja Model IndoBERT

Pengujian model *deep learning* IndoBERT dilakukan pada proporsi 20% data uji atau 1.748 ulasan guna memastikan komparasi performa yang objektif terhadap model probabilistik sebelumnya. Matriks hasil komputasi dari proses prediksi model divisualisasikan melalui *Confusion Matrix* pada Tabel 4.

Tabel 4. *Confusion Matrix* Model *IndoBERT*

	Predicted Negative	***Predicted Neutral***	***Predicted Positive***
Actual Negative	1070	29	97
Actual Neutral	19	8	32
Actual Positive	24	18	451

Kinerja agregat dari distribusi prediksi tersebut selanjutnya dikuantifikasi ke dalam metrik evaluasi pada Tabel 5.

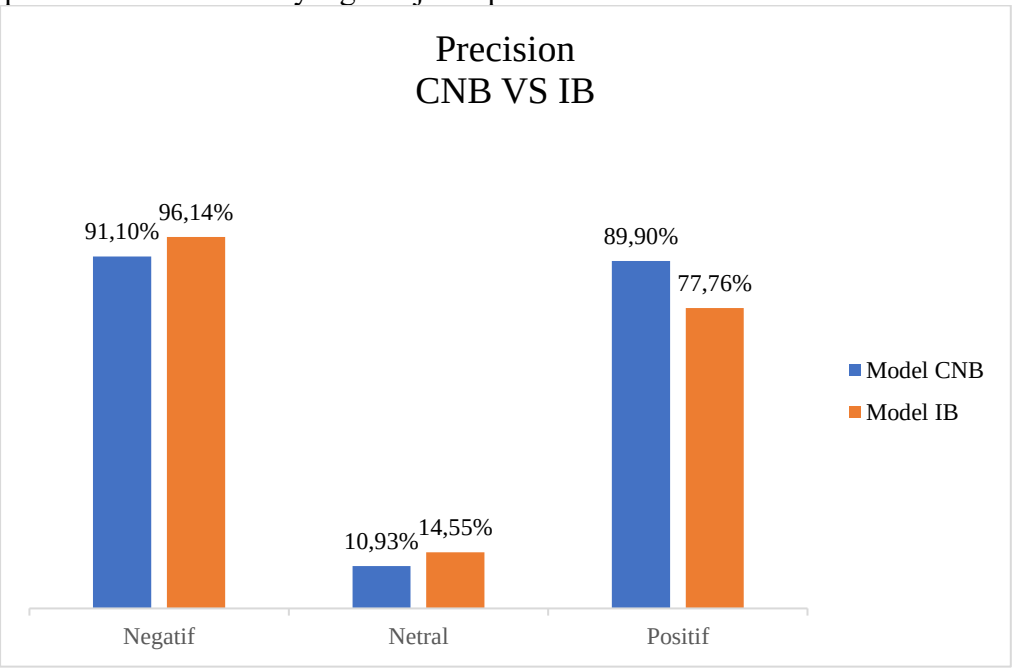
Tabel 5. Hasil Evaluasi Model IndoBERT

Sentimen	***Precision***	***Recall***	***F1-Score***
Negatif	96,14%	89,46%	92,68%
Netral	14,55%	13,56%	14,04%
Positif	77,76%	91,48%	84,06%
Rata-rata bobot	88,20%	87,47%	87,60%

Berdasarkan Tabel 5, IndoBERT memperoleh akurasi 87,47%. Pada kelas Negatif, model memperoleh *Precision* 96,14% dan *Recall* 89,46% (*F1-Score* 92,68%), yang menunjukkan kemampuan menangkap konteks keluhan operasional dengan jumlah *False Positive* yang rendah. Peningkatan terbesar terjadi pada kelas Positif, dengan *Recall* 91,48%, *Precision* 77,76%, dan *F1-Score* 84,06%. Hasil ini mengindikasikan kemampuan model mengurai kalimat apresiasi dan umpan balik meskipun ditulis dalam ragam informal dan tanpa penghapusan *stopwords*. Kelas Netral tetap menjadi kendala utama, dengan *Precision* 14,55%, *Recall* 13,56%, dan *F1-Score* 14,04%. Penerapan *Class Weights* pada fungsi kerugian (*loss function*) menekan lonjakan *False Positive* yang terjadi pada model dengan *oversampling*, tetapi ambiguitas linguistik ulasan Netral membuatnya sulit dipisahkan dari kelas bersentimen. Kondisi ini menurunkan *Macro F1* menjadi 63,59%, sementara *Weighted F1* tetap 87,60%.

3.4. Analisis Komparasi

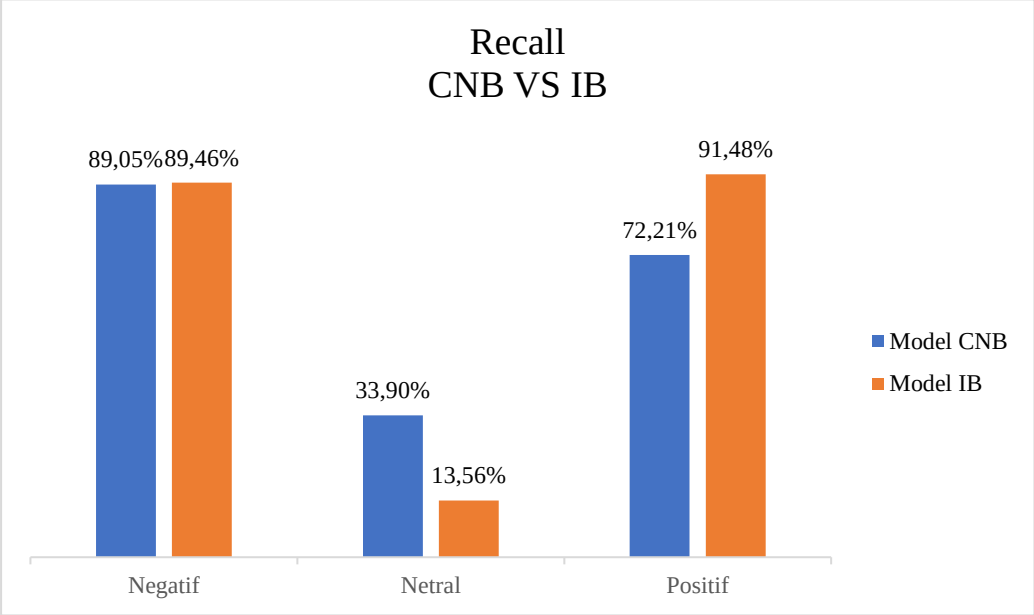
Evaluasi komparatif antara model probabilistik dan arsitektur *deep learning* dilakukan untuk menganalisis perbedaan performa pada teks informal dan data tidak seimbang. Perbandingan ditinjau per kelas sentimen sebelum diakumulasikan. Tinjauan pertama difokuskan pada metrik *Precision* yang disajikan pada Gambar 4.



Gambar 4. Grafik Komparasi Metrik *Precision*: CNB VS IB

Pada kelas Negatif, IndoBERT memperoleh *Precision* 96,14% dibandingkan CNB 91,10%, yang mengindikasikan mekanisme *self-attention* lebih efektif menekan *False Positive* pada ulasan keluhan. Pada kelas Positif, CNB memperoleh presisi lebih tinggi (89,90%) dibandingkan IndoBERT (77,76%), yang mengindikasikan sensitivitas pembobotan TF-IDF terhadap kosakata apresiasi tertentu. Pada kelas Netral, kedua model memperoleh presisi rendah (IndoBERT 14,55%; CNB 10,93%).

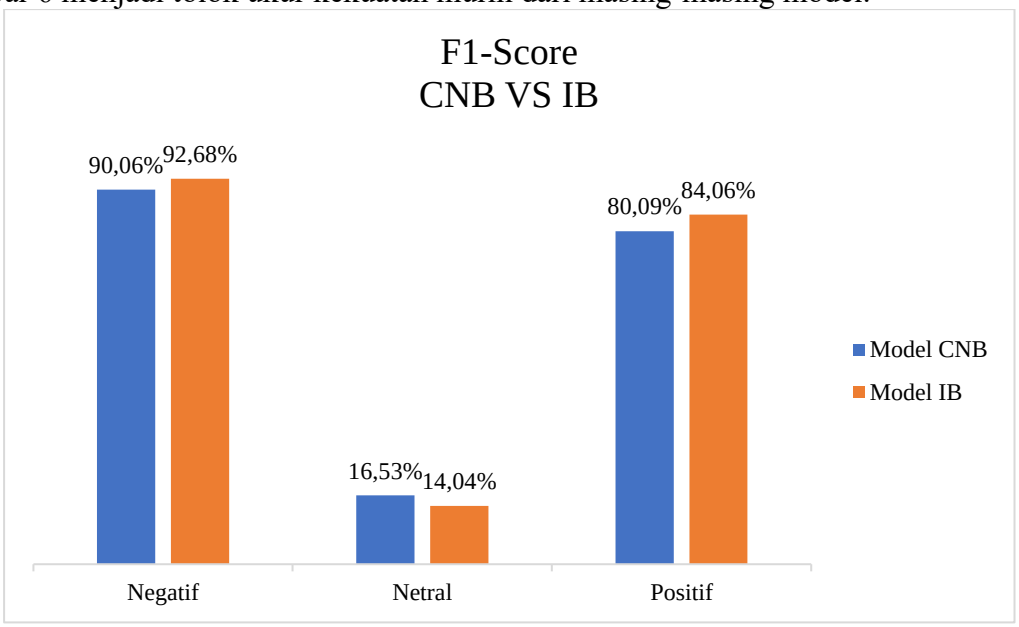
Kemampuan algoritma dalam menemukan kembali keseluruhan data aktual divisualisasikan melalui grafik *Recall* pada Gambar 5.



Gambar 5. Grafik Komparasi Metrik *Recall*: CNB VS IB

Perbedaan terbesar terdapat pada kelas Positif: IndoBERT memperoleh *Recall* 91,48%, sedangkan CNB 72,21%. Sebaliknya, pada kelas Netral CNB dengan SMOTE memperoleh *Recall* lebih tinggi (33,90%) dibandingkan IndoBERT (13,56%). Penambahan sampel sintetis memperluas batas keputusan CNB pada kelas Netral, meskipun disertai penurunan presisi.

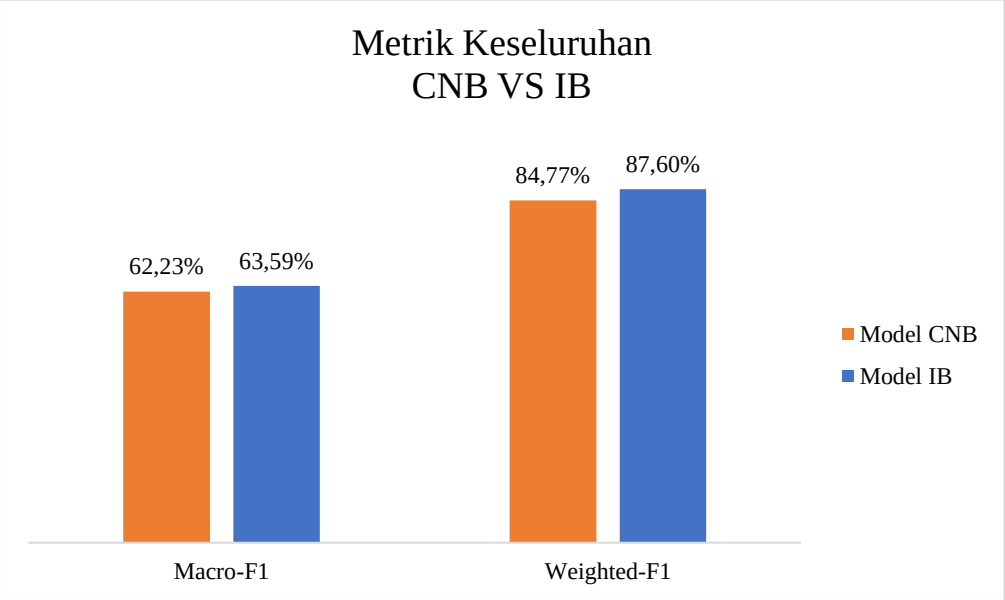
Sebagai parameter keseimbangan antara *Precision* dan *Recall*, grafik *F1-Score* per kelas pada Gambar 6 menjadi tolok ukur kekuatan murni dari masing-masing model.



Gambar 6. Grafik Komparasi Metrik F1-Score: CNB VS IB

IndoBERT memperoleh *F1-Score* lebih tinggi pada dua kelas utama, yaitu 92,68% (Negatif) dan 84,06% (Positif), dibandingkan CNB. Kestabilan ini mengindikasikan kemampuan arsitektur *Transformer* dalam menganalisis teks bahasa Indonesia informal.

Selanjutnya, performa kedua model diringkas dalam metrik agregat yang disajikan pada Gambar 7.



Gambar 7. Grafik Metrik Keseluruhan: CNB VS IB

IndoBERT memperoleh nilai lebih tinggi pada seluruh metrik agregat, dengan akurasi 87,47% atau 5,03 poin di atas CNB. Untuk menilai perlakuan terhadap data tidak seimbang, evaluasi ditinjau melalui *Macro F1-Score* dan *Weighted F1-Score*. IndoBERT memperoleh *Macro F1* 63,59% dibandingkan CNB 62,23%; selisih sebesar 1,36 poin ini relatif kecil sehingga keunggulannya pada kelas minoritas perlu ditafsirkan secara hati-hati.

Pada metrik *Weighted F1-Score*, IndoBERT memperoleh 87,60% dibandingkan CNB 84,77%. Berdasarkan capaian agregat tersebut, IndoBERT lebih sesuai digunakan untuk mengekstraksi sentimen dan keluhan operasional pada data ulasan dengan distribusi kelas yang tidak merata.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa distribusi opini mitra pengemudi pada kolom ulasan didominasi sentimen negatif yang berpusat pada kendala operasional, seperti regulasi tarif dan kebijakan suspensi. Penelitian ini juga menemukan ketidaksesuaian polaritas berupa *fake rating*, yaitu skor bintang tinggi yang disertai narasi keluhan. Koreksi berbasis leksikon (*Lexicon-Based Auto-Correction*) diterapkan untuk mengurangi ketidaksesuaian tersebut sebelum pemodelan.

Pada tahap evaluasi, IndoBERT memperoleh performa prediktif yang lebih baik dibandingkan Complement Naïve Bayes (CNB), dengan akurasi 87,47% dan *Weighted F1-Score* 87,60% berbanding 82,44% dan 84,77%. Keunggulan ini dapat dikaitkan dengan kemampuan IndoBERT mempertahankan konteks dua arah (*bidirectional*) tanpa penghapusan kata hubung, sehingga lebih sesuai untuk teks informal. Pada penanganan ketidakseimbangan kelas, *Class Weights* memberikan *Macro F1* 63,59% dibandingkan 62,23% pada SMOTE. Selisih tersebut kecil dan belum diuji signifikansinya, sehingga perbandingan ini bersifat indikatif.

Secara praktis, temuan ini dapat digunakan oleh pengelola platform sebagai rujukan evaluasi kebijakan operasional berbasis data (*data-driven*). Perbaikan pada transparansi skema bagi hasil dan respons terhadap keluhan teknis berpotensi menjaga keberlanjutan hubungan kemitraan dengan mitra pengemudi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Terbuka yang telah memfasilitasi penyusunan artikel ini, serta kepada seluruh pihak yang membantu selama proses pengumpulan data. Penulis juga menyampaikan terima kasih kepada keluarga atas dukungan yang diberikan selama penelitian berlangsung.

DAFTAR PUSTAKA

Andhika, F. R., Witanti, W., & Sabrina, P. N. (2025). Analisis Sentimen Menggunakan Metode IndoBERT pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe. *METIK Jurnal, 9*(2), 439–448. https://doi.org/10.47002/metik.v9i2.1098

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. *PETIR, 15*(2), 264–275. https://doi.org/10.33322/petir.v15i2.1733

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Destika, F., Prastika, R. A., Rahmaya, P. R., Christy, J., Hufad, A., & Achdiani, Y. (2026). Analisis Sistem Potongan Pendapatan dan Relasi Kerja Driver Ojek Online dalam Perspektif Materialisme Historis Karl Marx. *RISOMA: Jurnal Riset Sosial Humaniora dan Pendidikan, 4*(4), 214–228. https://doi.org/10.62383/risoma.v4i4.1737

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kusnaya, W. A., Cahyana, Y., & Juwita, A. R. (2025). Penerapan Metode Naive Bayes Multinomial dan Complement dalam Membandingkan Tingkat Akurasi terhadap Analisis Sentimen Kurikulum Merdeka. *Scientific Student Journal for Information, Technology and Science, 6*, 62–69. https://garuda.kemdiktisaintek.go.id/documents/detail/4870259

Madabushi, H. T., Kochkina, E., & Castelle, M. (2019). Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (pp. 125–134). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5018

Randhika, M. N., Young, J. C., Suryadibrata, A., & Mandala, H. (2021). Implementasi Algoritma Complement dan Multinomial Naïve Bayes Classifier pada Klasifikasi Kategori Berita Media Online. *Ultimatics: Jurnal Teknik Informatika, 13*(1), 19–25. https://doi.org/10.31937/ti.v13i1.1921

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)* (pp. 616–623).

Riyana, S. (2026). Analisis sentimen ulasan pengguna aplikasi Lalamove menggunakan Naïve Bayes, Support Vector Machine, dan Random Forest. *JTIK (Jurnal Teknik Informatika*

Kaputama), 10(1), 30–38. https://doi.org/10.59697/jtik.v10i1.1194

Salsabila, S. A., Priyatna, B., Hananto, A., & Tukino. (2025). Komparasi Kinerja Model Naive Bayes, SVM, dan BERT dalam Klasifikasi Sentimen Ulasan pada Aplikasi YUMMY. *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer, 4*(2), 42–47. https://doi.org/10.55123/storage.v4i2.5120

Sani, R. R., Pratiwi, Y. A., Winarno, S., Udayanti, E. D., & Alzami, F. (2022). Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia. *Jurnal Masyarakat Informatika, 13*(2), 85–98. https://doi.org/10.14710/jmasif.13.2.47983

Setiawan, E., Sartono, B., & Notodiputro, K. A. (2025). SMOTE and weighted random forest for classification of areas based on health problems in Java. *Journal of Applied Informatics and Computing, 9*(4), 1587–1592. https://doi.org/10.30871/jaic.v9i4.9933

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *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* (pp. 843–857). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.85