Teknika 19 (3): 947-959

JURNAL TEKNIKA
ISSN: 0854-3143
e-ISSN: 2622-3481
Journal homepage: http://jurnal.polsri.ac.id/index.php/teknika
Journal Email: teknika@polsri.ac.id

# Optimasi *Model* Hybrid *Random Forest* dan *Gradient Boosting* dalam Sentimen Ulasan Pengguna Shopee

**Sri Hartati*[1], Muhajir Arafat[2], Rusidi[3]**

[*1]Jurusan Manajemen Informatika, Universitas Mahakarya Asia, Baturaja, Indonesia
[2,3]Jurusan Teknik Informatika, Universitas Mahakarya Asia, Baturaja, Indonesia
*Email Penulis Korespondensi: hartatiakmi1984@gmail.com

***Abstrak***

*Dalam era transformasi digital, penggunaan e-commerce telah melekat erat dalam keseharian masyarakat Indonesia, dengan aplikasi Shopee sebagai salah satu platform terpopuler. Ulasan pengguna yang melimpah menawarkan peluang besar untuk memahami persepsi konsumen, namun analisis sentimen pada teks berbahasa Indonesia masih menghadapi tantangan karena sifatnya yang informal dan tidak seimbang. Penelitian ini bertujuan mengoptimalkan klasifikasi sentimen ulasan Shopee melalui pendekatan hybrid stacking yang menggabungkan Random Forest(RF) dan Gradient Boosting (GB) dengan Logistic Regression sebagai meta-learner. Dataset ulasan Shopee diproses menggunakan teknik preprocessing teks dan ekstraksi fitur TF-IDF. Evaluasi performa dilakukan dengan metrik akurasi, precision, recall, dan F1-score. Hasil menunjukkan bahwa model Hybrid Stacking memiliki kinerja paling unggul dengan akurasi 0,82, precision 0,82, recall 0,64, dan F1-score 0,65. Sementara model Random Forest mencatatkan akurasi 0,80, precision 0,86, recall 0,59, dan F1-score 0,59. Adapun model Gradient Boosting memperoleh akurasi 0,77, precision 0,62, recall 0,57, dan F1-score 0,55. Hasil ini menunjukkan bahwa pendekatan hybrid stacking mampu meningkatkan performa klasifikasi dibandingkan model individual. Temuan ini memberikan kontribusi terhadap pengembangan sistem analisis sentimen yang lebih akurat. Temuan ini menegaskan pentingnya pendekatan hybrid dalam analisis sentimen e-commerce berbahasa Indonesia, dan memberikan nilai tambah bagi pembangunan sistem pendukung keputusan yang lebih fleksibel dan akurat.*

***Kata kunci***—*Sentimen Analisis, Shopee, Random Forest, Gradient Boosting, Hybrid Stacking*

***Abstract***

*The period marked by digital transformation the use of e-commerce has become deeply integrated into the daily lives of Indonesian society, with Shopee emerging as among the most widely used platforms. The abundance of user reviews presents a significant opportunity to understand consumer perceptions; however, sentiment analysis on Indonesian-language text still faces challenges due to its informal nature and class imbalance. This study aims to optimize the sentiment classification of Shopee reviews through a hybrid stacking approach that combines Random Forest (RF) and Gradient Boosting (GB) with Logistic Regression as the meta-learner. The Shopee review dataset was processed using text preprocessing techniques and TF-IDF feature extraction. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results indicate that the Hybrid Stacking model achieved the best performance*

*Received: August 1, 2025; Online Available: September 3, 2025*

*with an accuracy of 0.82, precision of 0.82, recall of 0.64, and an F1-score of 0.65. In comparison, the Random Forest model obtained an accuracy of 0.80, precision of 0.86, recall of 0.59, and an F1-score of 0.59, while the Gradient Boosting model achieved an accuracy of 0.77, precision of 0.62, recall of 0.57, and an F1-score of 0.55. These findings demonstrate that the hybrid stacking approach can enhance classification performance compared to individual models. The findings of this study aid in building more accurate sentiment analysis systems and underscores the importance of hybrid approaches in Indonesian-language e-commerce sentiment analysis, providing added value for building more flexible and precise decision support systems.*

***Keywords—****Sentiment Analysis, Shopee, Random Forest, Gradient Boosting, Hybrid Stacking*

## 1. PENDAHULUAN

Dalam era transformasi digital, *platform e-commerce* telah menjadi bagian integral dari kegiatan masyarakat Indonesia. *Shopee*, sebagai salah satu pemain utama, berhasil menarik jutaan pengguna melalui fitur inovatif seperti *ShopeePay*, *flash sale*, dan sistem logistik terpadu [1]. Komentar pengguna dalam *platform* ini menyajikan wawasan penting untuk mengkaji pengalaman dan persepsi konsumen terhadap produk maupun layanan [2]. Namun, karakteristik ulasan berbahasa Indonesia yang cenderung informal, penuh singkatan, dan emotikon menghadirkan tantangan signifikan dalam *analisis sentimen* otomatis [3].

*Analisis sentimen* berperan sebagai alat strategis untuk mendukung pengambilan keputusan bisnis dan meningkatkan kualitas layanan. Metode tradisional seperti *NaiveBayes*, *LogisticRegression*, dan *Support Vector Machine* memang telah lama digunakan karena kemudahannya dalam implementasi dan interpretasi. Namun, penelitian di Indonesia menunjukkan bahwa ketika metode ini diterapkan pada *dataset* besar dan tidak seimbang, performanya cenderung menurun, di mana *Logistic Regression* memperoleh akurasi 77 %, SVM 76 %, dan *Naive Bayes* 70 % pada *analisis sentimen* imigrasi [4]. Temuan serupa juga ditemukan dalam studi internasional menggunakan *dataset* besar *Twitter US Airlines*, yang memperlihatkan bahwa meskipun *Logistic Regression* dan SVM masih lebih unggul dari pada *Naive Bayes*, efektivitas ketiganya tetap terbatas pada data yang kompleks dan tidak seimbang [5]. Hal ini menegaskan perlunya pendekatan yang lebih adaptif, seperti teknik *oversampling* atau pemanfaatan *model* berbasis *deep learning*, agar hasil analisis mencapai tingkat lebih akurat dan representatif. Maka dari itu diperlukan algoritma *ensemble* yang lebih tangguh.

Algoritma yang tangguh antara lain *Random Forest* dan *Gradient Boosting.* Alasan digunakannya kedua algoritma ini dikarenakan keunggulannya. Ketika dibandingkan dengan *Logistic Regression / SVM / Naïve Bayes* yang memiliki Model linear/naif tidak fleksibel dalam menangani non-linearitas dan interaksi fitur tinggi, *Random Forest* dan *Gradient Boosting* secara inheren menangani interaksi dan *non-linearitas* tanp*a eksplisit difiturisasi.* Selain itu Ketika dibandingkan dengan *AdaBoost / Extra Trees* yang rentan terhadap noise, *Extra Trees* terlalu acak maka *Gradient Boosting* punya regularisasi kuat dan *Random Forest* stabil tanpa *tuning intensif.* Dalam penelitian di Jurnal Teknika, *Random Forest* dan *Gradient Boosting* telah terbukti efektif dalam klasifikasi dibandingkan metode tradisional seperti *Naïve Bayes* maupun *SVM*.

*Random Forest* termasuk ke dalam kategori algoritma pembelajaran mesin yang berlandaskan teknik *ensemble* yang dikenal stabil dalam menangani data kompleks dan mampu meminimalkan *overfitting* melalui mekanisme *bagging* [6]. *Random Forest* menyediakan metrik seperti *feature importance*, model ini sulit diinterpretasikan secara menyeluruh karena terdiri dari banyak pohon keputusan, menjadikannya layaknya kotak hitam dalam pengambilan keputusan. Namun, penelitian terbaru memperlihatkan hasil bahwa model menunjukkan preferensi yang lebih tinggi terhadap kelas mayoritas dibandingkan kelas lainnya ketika diterapkan pada *dataset* dengan ketidakseimbangan yang tinggi, sehingga performa pada kelas minoritas menurun secara signifikan [7]. Hal ini sejalan dengan studi lain yang menegaskan bahwa pada *dataset* tidak seimbang, *Random Forest* lebih fokus pada kelas dominan sehingga menghasilkan performa yang tinggi di kelas mayoritas, tetapi lemah dalam memprediksi kelas minoritas [8]. *Random Forest*

(RF) dipilih karena *benchmark* besar-besaran menunjukkan RF umumnya menawarkan akurasi yang lebih tinggi dibandingkan *Logistic Regression* pada banyak dataset sekitar +2.9 % akurasi rata-rata [9].

*Gradient Boosting* merupakan algoritma *machine learning* berbasis *ensemble* yang unggul dalam meningkatkan akurasi karena pendekatan iteratifnya yang secara sistematis fokus pada residual *error* dari model sebelumnya. Dalam setiap tahap, algoritma menambahkan *weak learner* (biasanya pohon keputusan) yang dilatih untuk mengoreksi kesalahan prediksi sebelumnya, sehingga model akhir menjadi lebih akurat dan stabil [10]. Selain itu, proses ini juga menggunakan *pseudo-residuals* sebagai panduan menuju arah penurunan *loss*, serta mengatur *learning rate* secara hati-hati agar model tidak *overfit* menjadikannya sangat efektif dalam menangani masalah prediksi yang kompleks [11]. *Gradient Boosting* merupakan algoritma yang sangat bergantung pada penyetelan *hyperparameter* seperti *learning rate*, kedalaman pohon, dan jumlah iterasi (*n_estimators*) agar tidak mengalami *overfitting* dan tetap memberikan generalisasi yang baik [12]. Jika parameter tidak diatur secara hati-hati, struktur GBDT yang kompleks terutama dalam varian seperti *XGBoost, LightGBM*, atau *CatBoost* dapat menghasilkan model yang kuat di data latih namun gagal bekerja baik pada data baru. Oleh karena itu, diperlukan teknik regularisasi seperti *early stopping, subsampling*, atau penalti L1/L2 untuk menjaga kestabilan model dan mencegah *overfitting* [13].

Untuk mengatasi keterbatasan tersebut, pendekatan *hybrid stacking* kini semakin banyak diterapkan dalam *machine learning* modern. Penelitian terbaru menegaskan efektivitas *hybrid stacking* yang mengintegrasikan berbagai *base learner* seperti teknik pembelajaran mesin, kemudian disatukan oleh *meta-learner* untuk memaksimalkan akurasi serta stabilitas prediksi [14]. Pendekatan *stacking* berbasis *multimodal* (*teks + visual*) terbukti mampu meningkatkan performa klasifikasi ulasan *e-commerce* secara signifikan. Sebuah studi menggunakan model *BERT BiGRU Softmax* yang mengintegrasikan fitur teks dan citra, kemudian menerapkan strategi *stacking ensemble*, yang menghasilkan akurasi sebesar 90,9 % pada klasifikasi ulasan pelanggan [15]. Sementara itu, model *late fusion decision-level stacking*, yang menggabungkan *output* dari *BERT* (teks) dan *ResNet* (visual), juga menunjukkan peningkatan performa dibandingkan model *unimodal*, yang menegaskan efektivitas *stacking multimodal* dalam klasifikasi produk *e-commerce* [16].

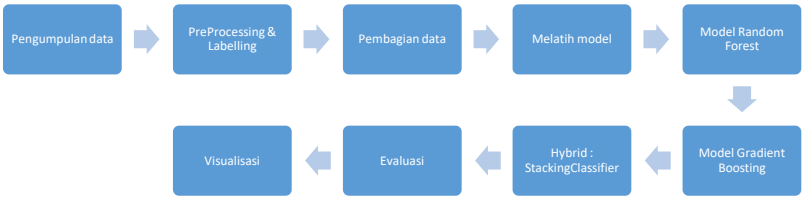
Meskipun berbagai pendekatan *ensemble* telah dikembangkan, kombinasi khusus RF dan GB dalam kerangka *stacking* untuk ulasan *Shopee* menggunakan bahasa negara kita masih sangat minim. Berdasarkan kondisi tersebut, penelitian ini mengembangkan model *hybrid stacking* RF–GB menggunakan TF-IDF untuk ekstraksi fitur dan *Logistic Regression* sebagai *meta-learner*. Metode ini ditargetkan dapat menghasilkan model yang memiliki ketahanan lebih baik sekaligus akurasi tinggi dalam *analisis sentimen Shopee*.

Berdasarkan pemaparan tersebut, masih terdapat kesenjangan penelitian dalam penerapan kombinasi khusus antara *Random Forest*dan *Gradient Boosting* untuk *analisis sentimen* ulasan pengguna *Shopee* berbahasa Indonesia, terutama pada skenario multi-kelas yang tidak seimbang. Penelitian sebelumnya lebih banyak berfokus pada penerapan metode tunggal atau kombinasi *multimodal* berbasis *deep learning*, sementara kajian yang secara eksplisit mengevaluasi efektivitas *hybrid stacking* RF–GB dengan representasi teks berbasis TF-IDF sangat minim keberadaanya. Sehubungan dengan hal tersebut, *novelty* penelitian ini terletak pada pengembangan model *hybrid stacking Random Forest*dan *Gradient Boosting* dengan *Logistic Regression* sebagai meta-learner, yang dirancang untuk mengoptimalkan akurasi dalam konteks bahasa Indonesia.

## 2. METODE PENELITIAN

Dalam penelitian sebelumnya metode yang digunakan mencakup proses *data preprocessing* untuk memastikan kualitas data sebelum dilakukan klasifikasi, di antaranya meliputi *data cleaning, data integration*, dan *data transformation*. Selanjutnya, diterapkan tiga algoritma pembelajaran mesin yaitu *C4.5, Naive Bayes*, dan *Random Forest* guna mengevaluasi

perbedaan performa dari setiap model dalam memprediksi [17]. Proses pengujian dilakukan menggunakan *cross-validation* untuk meminimalkan bias dan memastikan hasil yang lebih reliabel. Pendekatan kuantitatif eksperimental digunakan dalam penelitian ini guna membandingkan serta mengoptimalkan kinerja dua model, serta mengembangkan *model hybrid* berbasis *stacking* untuk klasifikasi sentimen ulasan pengguna aplikasi *Shopee*. Metode tersebut ditetapkan dengan pertimbangan perlunya pengukuran yang obyektif mengenai kemampuan tiap model dalam mengklasifikasikan sentimen teks berbahasa Indonesia. Penelitian ini menggunakan bantuan pemrograman Python dan pustaka *machine learning scikit-learn* di *Google Colab*. Seluruh tahapan dalam penelitian dilakukan secara sistematis seperti pada Gambar 1.




Gambar 1 Tahapan dalam penelitian

**Load *Dataset***

*Dataset* dari sumber sekunder yang diambil dari *Google Play Store* melalui proses *web scraping* menggunakan *library Python* khusus untuk pengambilan ulasan aplikasi. Data ini berasal dari aplikasi *Shopee*, salah satu *platform e-commerce* terpopuler di Indonesia, dengan total 10.000 entri ulasan pengguna. Setiap entri berisi informasi seperti teks ulasan (*content*), skor penilaian (*score*), nama pengguna (*userName*), versi aplikasi (*appVersion*), tanggal ulasan (at), dan beberapa metadata lainnya.

Hasil *Exploratory Data Analysis* (EDA) menunjukkan bahwa *dataset* memiliki 11 fitur dengan distribusi skor sebagai berikut: 5.391 ulasan positif, 3.788 ulasan negatif, dan 821 ulasan netral. Panjang rata-rata teks ulasan adalah sekitar 212 karakter. *Dataset* yang digunakan adalah hasil *scraping* ulasan pengguna dari aplikasi *Shopee*, yang disimpan dalam format CSV dengan nama *Shopee*10rb.csv. *Dataset* dimuat ke dalam memori menggunakan fungsi pd.read_csv() dari pustaka pandas. *Dataset* ini berisi dua kolom utama, yaitu:

1. Content : teks ulasan,
2. Score : skor rating numerik (1–5) dari pengguna.

***Preprocessing***

*Preprocessing* diterapkan guna memastikan data mentah berada dalam format yang sesuai untuk digunakan sebagai input *model*. Beberapa proses utama yang umum diterapkan meliputi:

1. *Text Cleaning*: menghilangkan tanda baca, angka, URL, dan karakter khusus yang tidak relevan.
2. *Case Folding*: normalisasi teks melalui perubahan semua huruf menjadi huruf kecil untuk menjaga keseragaman data.
3. *Stopword Removal*: penghapusan kata umum seperti “dan”, “yang”, “di”, yang tidak membawa makna penting.
4. *Wordcloud Visualization*: menampilkan representasi visual dari frekuensi kata dominan dalam data teks seperti tersaji pada Gambar 2.

Proses-proses ini dipandang penting karena dapat mengurangi kompleksitas data secara signifikan dan memperkuat akurasi model, khususnya dalam konteks *analisis sentimen* serta klasifikasi ulasan *e-commerce* [18], [19].




Gambar 2 *Wordcloud* dari ulasan aplikasi *Shopee*

***Labeling***

Labeling: skor rating dikonversi menjadi kategori sentimen seperti tersaji pada Gambar 3, sebagai berikut:

1. Skor 1–2 → Negatif
2. Skor 3 → Netral
3. Skor 4–5 → Positif

| sentimen | count |
|---|---|
| positif | 5391 |
| negatif | 3788 |
| netral | 821 |

Gambar 3 Konversi Skor ke Kategori Sentimen

**Pembagian data**

Pasca *preprocessing*, data teks ditransformasikan menjadi vektor numerik menggunakan pendekatan TF-IDF (*Term Frequency–Inverse Document Frequency*) yang mengukur ukuran relevansi suatu kata dalam dokumen dibandingkan dengan seluruh kumpulan teks. Pada tahap berikutnya, dataset dipisahkan dan dibagi menjadi:

1. Data latih (*training*) sebesar 80%
2. Data uji (*testing*) sebesar 20%

Pembagian dilakukan menggunakan metode *stratified* sampling agar distribusi label sentimen tetap seimbang pada kedua subset. Rumus (1) merupakan perhitungan TF-IDF (*Term Frequency-Inverse Document Frequency*).

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \log\left(\frac{n}{\mathrm{DF}(t)}\right) \quad (1)$$

Keterangan:

$t$ : kata (*term*)
$d$ : dokumen
$N$ : jumlah total dokumen
DF($t$) : jumlah dokumen yang mengandung kata $t$
TF($t$,$d$) : intensitas kemunculan kata $t$ pada teks dokumen $d$

***Training* dan *Testing***

1. Model Individual: *Random Forest*dan *Gradient Boosting*.

Kedua model dilatih menggunakan data latih, lalu diuji performanya menggunakan data uji. Model baseline yang digunakan adalah:

a. *Random Forest*: metode ensembel berbasis *bagging*, dengan banyak pohon keputusan yang dilatih secara paralel dan hasil voting digunakan untuk klasifikasi akhir.

b. *Gradient Boosting*: metode ensembel berbasis *Boosting*, di mana model dilatih secara bertahap dengan fokus pada memperbaiki kesalahan dari model sebelumnya.
2. Model Hybrid: *Stacking Classifier*

Untuk meningkatkan performa, dikembangkan model *hybrid* menggunakan pendekatan *stacking*, yaitu:

a. *Base learner*s: *Random Forest*dan *Gradient Boosting*
b. Meta-learner: *Logistic Regression*
c. Konfigurasi: *passthrough*=True dan 5-*fold cross-validation*
d. *Stacking Classifier* akan mengambil prediksi dari dua model dasar sebagai input tambahan untuk *meta-classifier*, sehingga diharapkan dapat memanfaatkan kekuatan masing-masing model untuk meningkatkan akurasi dan generalisasi.

**Evaluasi**

Tahapan evaluasi dan visualisasi dalam Fokus utama penelitian ini ialah mengevaluasi dan membandingkan performa model individual dalam klasifikasi. (*Random Forest* dan *Gradient Boosting*) serta model *hybrid* (*Stacking Classifier*) dalam pengklasifikasian sentimen ulasan pengguna aplikasi *Shopee*. Evaluasi dilakukan berdasarkan empat metrik utama yang menjadi praktik standar dalam klasifikasi multi-kelas:

1. *Accuracy*

Mengukur proporsi prediksi yang benar terhadap seluruh jumlah data uji seperti tersaji pada Rumus (2).

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2)$$

Keterangan:
*TP (True Positive):* Jumlah data positif yang berhasil diprediksi dengan benar sebagai positif.
*TN (True Negative):* Jumlah data negatif yang berhasil diprediksi dengan benar sebagai negatif.
*FP (False Positive):* Jumlah data negatif yang salah diprediksi sebagai positif (kesalahan tipe I).
*FN (False Negative):* Jumlah data positif yang salah diprediksi sebagai negatif (kesalahan tipe II).

2. *Precision* (*Macro Average*)

Menilai ketepatan prediksi positif dari setiap kelas, kemudian dirata-ratakan. Metrik ini penting untuk mengetahui seberapa akurat model dalam menghindari false positive, terutama pada kelas minoritas seperti netral atau negatif seperti tersaji pada Rumus (3).

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)$$

Keterangan:
*TP (True Positive):* Jumlah data positif yang berhasil diprediksi dengan benar sebagai positif.
*FP (False Positive):* Jumlah data negatif yang salah diprediksi sebagai positif (kesalahan tipe I).

3. *Recall* (*Macro Average*)

Mengukur kemampuan model dalam mendeteksi seluruh *instance* positif dari masing-masing kelas seperti tersaji pada Rumus (4). Metrik ini penting dalam konteks ketidakseimbangan kelas (*class imbalance*), agar model tidak bias terhadap kelas mayoritas.

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (4)$$

Keterangan:
*TP (True Positive):* Jumlah data positif yang berhasil diprediksi dengan benar sebagai positif.
*FN (False Negative):* Jumlah data positif yang salah diprediksi sebagai negatif (kesalahan tipe II).

4. F1-Score (*Macro Average*)

Dengan menggunakan *harmonic* mean dari *precision* dan *recall*, *F1-score* memberikan gambaran evaluatif yang menyeimbangkan keakuratan prediksi dan kelengkapan deteksi model.

$$\text{F1 Score } = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

**Visualisasi**

Selain evaluasi numerik, penelitian ini juga menggunakan visualisasi untuk memperjelas interpretasi hasil, antara lain:

1. *Confusion Matrix* dalam bentuk *heatmap* untuk melihat pola kesalahan prediksi antar kelas.
2. *Bar chart* visualisasi metrik dari semua model untuk mempermudah perbandingan.

## 3. HASIL DAN PEMBAHASAN

Studi ini mengevaluasi performa individu dan membandingkan performa *Random Forest*, *Gradient Boosting*, serta model *Hybrid stacking Classifier* dalam pengklasifikasian sentimen ulasan pengguna aplikasi *Shopee*. Pengukuran efektivitas *model* dilakukan berdasarkan metrik evaluasi yang dihitung menggunakan data uji sebanyak 80% dari total *dataset* yang telah diolah melalui langkah-langkah *preprocessing* dan vectorisasi TF-IDF. Bagian ini menampilkan hasil evaluasi kuantitatif dan pembahasan kualitatif untuk melengkapi hasil evaluasi tersebut. Penelitian ini membandingkan kinerja model yaitu Random Forest, *Gradient Boosting*, dan model *Hybrid stacking*Classifier.

**Model *Random Forest***

Hasil pengujian model *Random Forest* telah memperlihatkan bahwa model ini menghasilkan akurasi keseluruhan 80%, disertai presisi paling tinggi pada kategori netral (1,00) namun diiringi dengan *recall* yang sangat rendah (0,07). Hal ini mengindikasikan bahwa meskipun model sangat yakin ketika memprediksi kelas netral, namun sangat jarang berhasil mengenali data netral dengan benar, sehingga nilai *F1-score* untuk kelas netral hanya 0,13. Sebaliknya, model menunjukkan performa yang baik dan relatif seimbang pada kelas negatif (*precision* 0,73, *recall* 0,83) dan positif (*precision* 0,85, *recall* 0,89), yang ditunjukkan dengan *F1-score* yang dicatat pada tiap kelas mencapai 0,78 dan 0,87. Nilai macro *average recall* sebesar 0,60 dan *F1-score* 0,59 mengindikasikan bahwa model ini cenderung bias terhadap kelas mayoritas (negatif dan positif), sementara kelas minoritas (netral) kurang terdeteksi dengan baik. Hal ini memperkuat pentingnya pendekatan hybrid atau balancing agar model tidak mengabaikan prediksi kelas dengan jumlah data yang relatif kecil dengan sebaran seperti pada Tabel 1.

Tabel 1 *Classification Report Random Forest*

| | ***precision*** | ***recall*** | ***f1-score*** | ***support*** |
|---|---|---|---|---|
| **negatif** | 0,73 | 0,83 | 0,78 | 758 |
| **netral** | 1,00 | 0,07 | 0,13 | 164 |
| **positif** | 0,85 | 0,89 | 0,87 | 1078 |
| **accuracy** | | | 0,80 | 2000 |
| **macro avg** | 0,86 | 0,60 | 0,59 | 2000 |
| **weighted avg** | 0,82 | 0,80 | 0,77 | 2000 |

***Gradient Boosting***

Model *Gradient Boosting* menunjukkan performa klasifikasi yang cukup baik secara umum, dengan akurasi sebesar 77% dan *F1-score weighted average* sebesar 0,75. Model ini bekerja cukup baik pada kelas positif, dengan precision sebesar 0,83, *recall* 0,86, dan *F1-score* 0,84. Temuan ini mengindikasikan bahwa *model* memiliki kemampuan yang tinggi dalam mengidentifikasi sentimen positif. Pada kelas negatif, performa juga relatif kuat, dengan *precision*

0,71 dan *recall* 0,81, menghasilkan *F1-score* sebesar 0,76. Namun, performa menurun drastis pada kelas netral, di mana *precision* hanya mencapai 0,33 dan *recall* sangat rendah di angka 0,03. Nilai *F1-score* pada kelas ini hanya sebesar 0,06, menandakan bahwa model hampir tidak mampu mengidentifikasi sentimen netral dengan benar. Nilai *macro average F1-score* sebesar 0,55 juga menunjukkan ketidakseimbangan dalam performa model terhadap seluruh kelas seperti tersaji pada Tabel 2. Secara keseluruhan, meskipun *Gradient Boosting* berhasil mengenali sentimen mayoritas dengan baik, kelemahannya dalam mengenali kelas minoritas seperti netral menjadi salah satu faktor yang membatasi akurasi dan generalisasi model, serta memperkuat urgensi pengembangan model hybrid atau strategi balancing dalam penelitian ini.

Tabel 2 Classification Report *Gradient Boosting*

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| **negatif** | 0,71 | 0,81 | 0,76 | 758 |
| **netral** | 0,33 | 0,03 | 0,06 | 164 |
| **positif** | 0,83 | 0,86 | 0,84 | 1078 |
| **accuracy** | - | - | 0,77 | 2000 |
| **macro avg** | 0,62 | 0,57 | 0,55 | 2000 |
| **weighted avg** | 0,74 | 0,77 | 0,75 | 2000 |

**Hybrid (*Stacking*)**

Model Hybrid (*Stacking*) yang menggabungkan *Random Forest* dan *Gradient Boosting* menunjukkan performa yang lebih unggul secara keseluruhan dibandingkan model individu. Akurasi yang dicapai sebesar 82%, dengan *weighted average F1-score* sebesar 0,80, menunjukkan tingkat prediksi yang baik secara umum. Model mampu mengklasifikasikan kelas positif dengan sangat baik, ditunjukkan oleh precision 0,87, *recall* 0,88, dan *F1-score* 0,87 seperti tersaji pada Tabel 3. Untuk kelas negatif, performa juga cukup kuat, dengan *precision* 0,75 dan *recall* 0,87, menghasilkan *F1-score* sebesar 0,80. Hal yang menarik, model ini juga mengalami peningkatan signifikan pada kelas netral, yang sebelumnya menjadi kelemahan utama pada model individu. Meskipun *recall* untuk kelas netral masih rendah (0,16), *precision* meningkat menjadi 0,84, sehingga *F1-score* naik menjadi 0,27. Nilai *macro average F1-score* sebesar 0,65 lebih tinggi dibandingkan *Random Forest* (0,59) dan *Gradient Boosting* (0,55), menunjukkan bahwa pendekatan *hybrid* lebih seimbang dalam memprediksi semua kelas. Hasil ini memperkuat bahwa strategi *stacking* mampu menggabungkan keunggulan masing-masing *model* dasar untuk meningkatkan akurasi dan kestabilan klasifikasi sentimen, terutama dalam menangani kelas minoritas seperti netral.

Tabel 3 Classification Report Hybrid (*Stacking*)

| | ***precision*** | ***recall*** | ***f1-score*** | ***support*** |
|---|---|---|---|---|
| **negatif** | 0,75 | 0,87 | 0,80 | 758 |
| **netral** | 0,84 | 0,16 | 0,27 | 164 |
| **positif** | 0,87 | 0,88 | 0,87 | 1078 |
| **accuracy** | - | - | 0,82 | 2000 |
| **macro avg** | 0,82 | 0,64 | 0,65 | 2000 |
| **weighted avg** | 0,82 | 0,82 | 0,80 | 2000 |

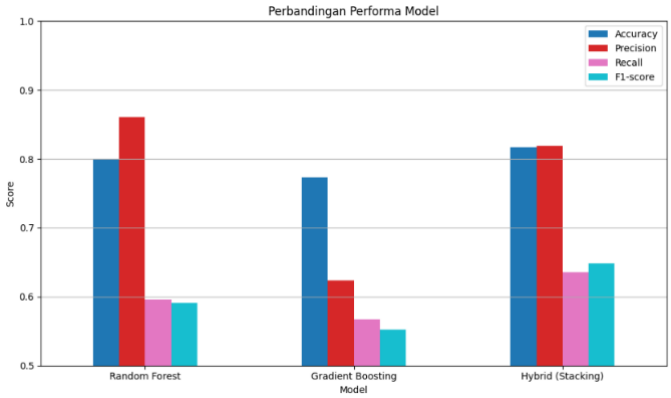
Penelitian memberikan hasil evaluasi kuantitatif yang merupakan hasil pengujian performa *model* dalam bentuk evaluasi seperti Tabel 4.

Tabel 4 Hasil Evaluasi Kinerja *Model*

| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1-Score*** |
|---|---|---|---|---|
| Random Forest | 0,80 | 0,86 | 0,59 | 0,59 |

| *Model* | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|
| *Gradient Boosting* | 0,77 | 0,62 | 0,57 | 0,55 |
| Hybrid (*Stacking*) | 0,82 | 0,82 | 0,64 | 0,65 |

Hasil evaluasi kuantitatif yang merupakan hasil pengujian performa *model* selain dibuat dalam bentuk tabel evaluasi, juga ditampilkan dalam bentuk grafik perbandingan kinerja *model* seperti tersaji pada Gambar 4.




Gambar 4 Grafik perbandingan kinerja *model*

Grafik perbandingan kinerja *model* menunjukkan bahwa *model* Hybrid (*Stacking*) unggul dalam semua metrik dibandingkan dengan kedua *model* individual. *Precision* menunjukkan peningkatan dari 0,62 hingga mencapai 0,82, sedangkan *F1-Score* naik dari 0,55 menjadi 0,65, menandakan bahwa pendekatan *stacking* berhasil meningkatkan keseimbangan antara ketepatan dan kelengkapan *model*.

**Pembahasan Kualitatif**

Penelitian ini selain menghasilkan evaluasi kuantitatif juga dilengkapi pembahasan. Pembahasan ini disajikan secara kualitatif guna memberikan interpretasi mendalam terhadap performa masing-masing *model*. Pendekatan ini bertujuan untuk menguraikan bagaimana karakteristik dan mekanisme kerja dari *model* Random Forest, *Gradient Boosting*, dan *Hybrid stacking* mempengaruhi hasil klasifikasi sentimen yang dihasilkan. Dengan menelaah kekuatan dan kelemahan tiap algoritma berdasarkan hasil pengujian, dapat diperoleh pemahaman yang lebih utuh mengenai validitas dan kinerja metode yang diterapkan dalam kajian *analisis sentimen* ulasan pengguna *Shopee*.

Merujuk pada evaluasi yang telah dilaksanakan, disimpulkan bahwa *Random Forest* mampu mencapai akurasi yang tinggi (0,80) dan *precision* yang sangat baik (0,86), namun *recall*-nya masih rendah (0.59), yang berarti *model* ini masih kurang mampu mengenali seluruh *instance* dari kelas minoritas, terutama sentimen netral.

*Model Gradient Boosting* memiliki akurasi sedikit lebih rendah (0,77) dan *precision* (0,62), namun secara umum masih kompetitif, terutama karena metode ini bertahap membenahi kesalahan prediksi sebelumnya.

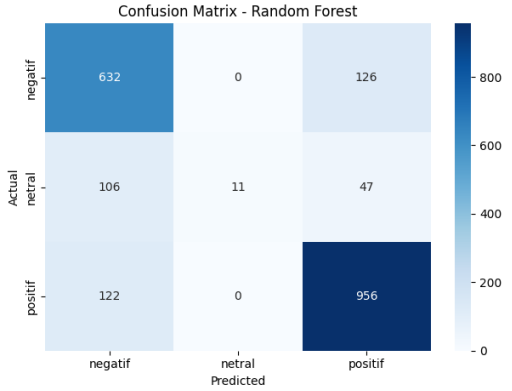
*Model Hybrid stacking* menggabungkan kedua pendekatan dengan *Logistic Regression* sebagai *meta-classifier*, dan hasilnya menunjukkan peningkatan signifikan dalam *recall* dan *F1-score*. Hal ini menunjukkan bahwa penggunaan *stacking* mampu menangkap pola klasifikasi yang tidak bisa ditangani optimal oleh masing-masing *model* secara individual.

**Analisis Kesalahan Klasifikasi (*Confusion Matrix*)**

Masing-masing *model* memiliki pola kesalahan klasifikasi yang berbeda terhadap tiga kelas sentimen, yaitu negatif, netral, dan positif. Analisis ini dilakukan untuk mengetahui seberapa besar kesalahan *model* dalam mengenali dan mengklasifikasikan setiap label sentimen.

**Confusion Matrix Random Forest**

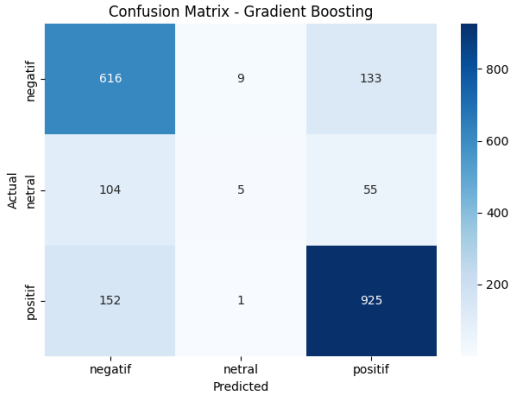
Pada *model Random Forest* prediksi terhadap kelas negatif dan positif relatif kuat dengan jumlah prediksi benar masing-masing sebesar 632 dan 956. Namun, *model* ini mengalami kesulitan signifikan dalam mengenali kelas netral, di mana hanya 11 *instance* yang berhasil diklasifikasikan dengan benar dari total 164. Sebagian besar data netral justru diklasifikasikan sebagai negatif (106 *instance*) atau positif (47 *instance*). Hal ini menunjukkan bahwa *Random Forest* cenderung mengabaikan kelas netral karena dominasi jumlah data positif dan negatif, serta kemungkinan karena distribusi data yang tidak seimbang seperti tersaji pada Gambar 5.




Gambar 5 *Confusion Matrix Random Forest*

**Confusion Matrix *Gradient Boosting***

Pada *model Gradient Boosting* pola kesalahan tidak jauh berbeda. *Model* mampu mengenali kelas positif dengan cukup baik (925 benar), namun kembali menunjukkan performa yang sangat rendah pada kelas netral, di mana hanya 5 instance yang diprediksi benar. Sebanyak 104 *instance* netral diprediksi sebagai negatif, dan 55 sebagai positif, mengindikasikan bahwa *model Gradient Boosting* belum mampu menangani ketidakseimbangan kelas secara efektif. Bahkan lebih buruk dari *Random Forest*dalam mengenali kelas netral seperti tersaji pada Gambar 6.

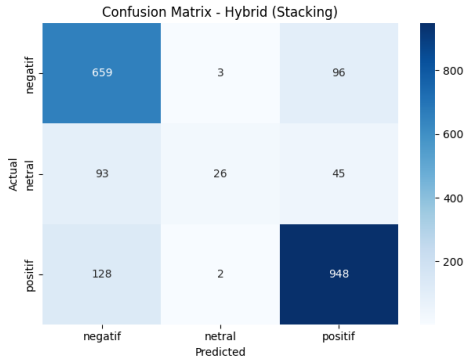



Gambar 6 *Confusion Matrix Gradient Boosting*

**Confusion Matrix Hybrid (*Stacking*)**

*Model* Hybrid (*Stacking*) menunjukkan performa yang lebih seimbang dan akurat. *Model* ini berhasil mengklasifikasikan 26 *instance* kelas netral dengan benar, jumlah yang jauh lebih tinggi dibandingkan dua *model* sebelumnya. Selain itu, prediksi benar pada kelas negatif mencapai 659 *instance*, dan pada kelas positif sebanyak 948 *instance*, hanya sedikit di bawah performa terbaik *model* individu. Kesalahan klasifikasi pada kelas netral juga lebih terdistribusi secara wajar, dengan 93 *instance* salah diklasifikasikan sebagai negatif dan 45 sebagai positif,

menunjukkan bahwa *model* hybrid lebih sensitif terhadap pola minoritas dalam data seperti tersaji pada Gambar 7.




Gambar 7 *Confusion Matrix Hybrid* (*Stacking*)

Berdasarkan hasil evaluasi, model *Hybrid Stacking* yang menggabungkan *Random Forest* (RF) dan *Gradient Boosting* (GB) dengan *Logistic Regression* sebagai *meta learner* menunjukkan performa yang lebih unggul dibandingkan model individual. Keunggulan ini dikarenakan beberapa faktor:

1. Kombinasi Kekuatan Dua *Algoritma Ensemble*

   RF unggul dalam mengurangi *overfitting* melalui teknik *bagging*, sementara GB efektif meningkatkan akurasi melalui pembelajaran iteratif yang fokus pada kesalahan prediksi sebelumnya (*boosting*). Berbeda dengan model linear seperti *Logistic Regression* atau SVM yang memiliki keterbatasan pada interaksi non-linear, RF dan GB dapat secara alami menangkap pola interaksi antar fitur yang kompleks.
2. Peran *Meta Learner* dalam *Stacking*

   *Logistic Regression* sebagai *meta learner* berfungsi menggabungkan output probabilistik dari RF dan GB, memilih pola prediksi terbaik, dan meminimalkan bias masing-masing *base learner*.
3. Pengurangan Varians dan Bias

   RF cenderung memiliki varians rendah namun bias moderat, sementara GB memiliki bias rendah namun lebih rentan terhadap varians tinggi jika parameter tidak diatur dengan tepat.
4. Kinerja Konsisten pada Kelas Minoritas

   Hasil pengujian menunjukkan bahwa *hybrid stacking* mampu meningkatkan *recall* pada kelas minoritas dibandingkan model individual.

## 4. KESIMPULAN

Merujuk pada temuan yang diperoleh dari pengujian dan analisis atas *model Random Forest*, *Gradient Boosting*, dan *Hybrid stacking Classifier* untuk klasifikasi sentimen ulasan pengguna aplikasi *Shopee*, maka dapat disimpulkan hal-hal berikut:

1. *Model* Hybrid (*Stacking Classifier*) menunjukkan performa terbaik secara keseluruhan, dengan nilai akurasi sebesar 82%, precision 0,82 kemudian *recall* 0,64 dan *F1-score* 0,65. Angka ini lebih tinggi dibandingkan *model Random Forest*(F1-score: 0,59) dan *Gradient Boosting* (*F1-score*: 0,55).
2. Penggunaan *stacking* dengan *meta-classifier Logistic Regression* dan *passthrough True* berhasil memadukan kekuatan dua *model* dasar, yaitu *Random Forest* (yang unggul dalam stabilitas klasifikasi mayoritas) dan *Gradient Boosting* (yang unggul dalam koreksi kesalahan), sehingga menghasilkan *model* yang lebih seimbang dan akurat.
3. Berdasarkan analisis *confusion matrix*, *model hybrid* juga mampu meningkatkan kemampuan deteksi terhadap kelas minoritas, terutama kelas netral, dengan jumlah prediksi benar meningkat dari hanya 11 (RF) dan 5 (GB) menjadi 26 instance (*Hybrid*).

4. Kelebihan *model hybrid* terletak pada kestabilannya dalam klasifikasi multi-kelas, serta kemampuannya mengurangi bias terhadap kelas mayoritas. *Model* ini juga lebih adaptif dalam menangani distribusi data yang tidak seimbang.
5. Kekurangan yang masih ada pada *model hybrid* adalah *recall* yang masih rendah pada kelas netral (0,16), menunjukkan bahwa meskipun terjadi perbaikan, *model* masih memiliki keterbatasan dalam mengenali opini yang bersifat netral atau ambigu secara linguistik.

## 5. SARAN

Rekomendasi pengembangan riset berikutnya untuk mengatasi keterbatasan penelitian ini, antara lain sebagai berikut:

1. Perlu dilakukan *balancing* data secara eksplisit, seperti menggunakan *SMOTE* atau teknik resampling lainnya, untuk meningkatkan *recall* pada kelas minoritas terutama sentimen netral.
2. Eksplorasi model *meta-classifier* yang lebih kompleks, seperti *Support Vector Machine* atau *Gradient Boosting* sebagai meta-learner, berpotensi memberikan peningkatan performa lebih lanjut dalam arsitektur *stacking*.
3. Penelitian selanjutnya dapat mengintegrasikan metode *embedding* seperti *Word2Vec*, *FastText*, atau *BERT*, agar representasi fitur teks menjadi lebih semantik dan kontekstual dibanding *TF-IDF*.
4. Evaluasi dapat diperluas pada domain lain, seperti ulasan layanan publik, perbankan, atau transportasi digital, agar model lebih general dan dapat digunakan secara praktis di sektor industri yang lebih luas.
5. Penerapan model dapat dikembangkan menjadi sistem *real-time sentiment* monitoring untuk *Shopee* atau platform *e-commerce* lainnya sebagai bagian dari sistem rekomendasi dan peningkatan layanan pelanggan.

## UCAPAN TERIMA KASIH

Dengan penuh rasa syukur, penulis mengucapkan terima kasih kepada Tim Redaksi Jurnal Teknika Politeknik Negeri Sriwijaya yang telah mendukung, memfasilitasi dan memberikan kesempatan untuk penerbitan artikel ini.

## DAFTAR PUSTAKA

[1] T. Dewi, A. Asrianda, and Y. Afrillia, “Sentiment Analysis of Customer Satisfaction Towards Shopee and Lazada E-commerce Platform Using the Random Forest Algorithm Classifier,” *International Journal of Engineering, Science and Information Technology*, vol. 5, no. 1, pp. 229–235, Dec. 2024, doi: 10.52088/ijesty.v5i1.692.

[2] M. Rusdi Rahman, A. Febri Diansyah, and H. Hanafi, “Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method,” *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 20–24, Nov. 2023, doi: 10.25139/inform.v9i1.5465.

[3] J. Y. B. Yin, N. H. M. Saad, and Z. Yaacob, “Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee,” *TEM Journal*, vol. 11, no. 4, pp. 1508–1519, Nov. 2022, doi: 10.18421/TEM114-11.

[4] P. Assiroj, A. Kurnia, and S. Alam, “The performance of Naïve Bayes, support vector machine, and logistic regression on Indonesia immigration sentiment analysis,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3843–3852, Dec. 2023, doi: 10.11591/eei.v12i6.5688.

[5] Md. T. H. K. Tusar and Md. T. Islam, “A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data,” Oct. 2021.

[6] R. Jayadi, "Sentiment Analysis Of Indonesian E-Commerce Product Reviews Using Support Vector Machine Based Term Frequency Inverse Document Frequency 1 Siti Fidyanti Nurfadila Hadju," *J Theor Appl Inf Technol*, vol. 15, p. 17, 2021.

[7] L. Dube and T. Verster, "Interpretability of the random forest model under class imbalance," *Data Science in Finance and Economics*, vol. 4, no. 3, pp. 446–468, 2024, doi: 10.3934/dsfe.2024019.

[8] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies (Basel)*, vol. 13, no. 3, Mar. 2025, doi: 10.3390/technologies13030088.

[9] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, Jul. 2018, doi: 10.1186/s12859-018-2264-5.

[10] L. W. Rizkallah, "Enhancing the performance of gradient boosting trees on regression problems," *J Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-025-01071-3.

[11] A. Eisenbürger, D. Otten, A. Hudde, and F. Hopfgartner, "Training Gradient Boosted Decision Trees on Tabular Data Containing Label Noise for Classification Tasks," Jan. 2025.

[12] P. Florek and A. Zagdański, "Benchmarking state-of-the-art gradient boosting algorithms for classification," May 2023.

[13] S. Cui, A. Sudjianto, A. Zhang, and R. Li, "Enhancing Robustness of Gradient-Boosted Decision Trees through One-Hot Encoding and Regularization," May 2023.

[14] L. Liu, G. Zhao, W. Liang, and Z. Jian, "Hybrid stacking ensemble algorithm and simulated annealing optimization for stability evaluation of underground entry-type excavations," *Underground Space (new)*, vol. 17, pp. 25–44, Aug. 2024, doi: 10.1016/j.undsp.2023.11.002.

[15] J. Liu, Y. Sun, Y. Zhang, and C. Lu, "Research on Online Review Information Classification Based on Multimodal Deep Learning," *Applied Sciences (Switzerland)*, vol. 14, no. 9, May 2024, doi: 10.3390/app14093801.

[16] Bi Ye, Wang Shuo, and Fang Zhongrui, "A Multimodal Late Fusion Model for E-Commerce Product Classification-10b," *arxiv.org*, Aug. 2020.

[17] J. Teknika, "Teknika 17 (2): 419-426 Perbandingan Kinerja Algoritma C4.5. Naive Bayes Dan Random Forest Dalam Prediksi Penyakit Jantung," *IJCCS*, vol. x, No.x, pp. 1–5.

[18] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Applied Sciences (Switzerland)*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178765.

[19] M. Arief and M. B. M. Deris, "Text Preprocessing Impact for Sentiment Classification in Product Review," in *2021 6th International Conference on Informatics and Computing, ICIC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICIC54025.2021.9632884.