Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2790~2798
e-ISSN: 2548-6861

Hybrid Machine Learning for Knowledge Discovery in E-Commerce Reviews

Jeremiah Alwin Siahaan [1]*, Lailla Syal Syabilla [2], M. Thoriqul Fadli [3], Mei Intan Natasyah [4], Allsela Meiriza [5], Ken Ditha Tania [6], Ahmad Rifai [7]
* Sistem Informasi, Universitas Sriwijaya
09031182328007@student.unsri.ac.id[1], 09031382328125@student.unsri.ac.id[2], 09031282328037@student.unsri.ac.id [3], 09031282328047@student.unsri.ac.id[4], allsela@unsri.ac.id,[5] kenya.tania@gmail.com[6], ahmadrifai@ilkom.unsri.ac.id[7]

Article Info

Article history:

Received 2026-03-05
Revised 2026-05-18
Accepted 2026-05-25

Keyword:

K-Means Clustering,
Knowledge Discovery in Database,
Random Forest,
Tokopedia,
Uninformative Review.

ABSTRACT

The rapid growth of e-commerce platforms like Tokopedia has triggered a massive accumulation of over 65,000 customer reviews, yet it is often accompanied by information pollution in the form of non-informative reviews that hinder consumer decision-making processes. This research aims to extract new knowledge regarding these review characteristics through the implementation of the Knowledge Discovery in Database (KDD) framework, integrating a hybrid K-Means Clustering and Random Forest algorithm. Diverging from conventional classification approaches, this study utilizes K-Means as an exploratory instrument to naturally map six latent topic patterns of reviews based on their textual structure. Experiments were conducted on 35,000 data samples using TF-IDF features enriched by cluster labels as structural predictors. The results indicate that the hybrid model achieves 94.41% accuracy with an F1-score of 0.90 for the non-informative class, showing high stability via 5-Fold Cross-Validation (94.56% ± 0.19%) . The most crucial knowledge discovery is evidenced through SHAP analysis, where the cluster feature ranks 7th out of 1,001 predictor features, confirming that semantic grouping provides a richer structural context than pure lexical features . Furthermore, error analysis reveals specific linguistic challenges such as sarcasm and semantic ambiguity as constraints in automated review detection . This research provides a managerial contribution to e-commerce platforms in enhancing information quality and mitigating information overload issues.

I. PENDAHULUAN

Tokopedia merupakan salah satu situs *e-commerce* terkemuka di Indonesia, memberikan kesempatan kepada penggunanya untuk memberikan *feedback* mengenai produk setelah melakukan pembelian [1]. Salah satu fitur terpenting pada Tokopedia adalah fitur ulasan yang berguna sebagai instrumen krusial dalam membangun kepercayaan antara penjual dan pembeli [1]. Secara teoritis, ulasan pelanggan bertindak sebagai sumber informasi sekunder bagi calon pembeli dalam mengatasi risiko ketidakpastian saat proses pengambilan keputusan. Hal ini sejalan dengan penelitian Azimi dan Andonova [2] yang menyatakan bahwa ulasan daring dianggap sebagai salah satu sumber informasi terpenting yang mempengaruhi keputusan pembelian konsumen. Lebih lanjut, merujuk pada data yang dikutip oleh Azimi dan Andonova [2], tercatat bahwa lebih dari 93% konsumen menggunakan ulasan daring sebagai referensi untuk membeli produk. Mengingat pentingnya ulasan daring dalam pengambilan keputusan, maka sangat penting untuk memahami apa yang membuat ulasan tersebut bermanfaat dari sudut pandang konsumen.

Faktanya tidak semua ulasan yang diberikan pengguna bersifat informatif. Dalam penelitian ini, ulasan dikategorikan sebagai tidak informatif apabila memenuhi satu atau lebih kriteria berikut seperti jumlah kata yang terbatas pada ulasan mentah, seluruh konten terdiri dari simbol, karakter acak, atau emoji tanpa teks verbal, mengandung kata berulang tanpa variasi makna seperti “okokok” atau “bagusss” dan tidak mengandung informasi spesifik mengenai produk, layanan, maupun pengalaman pembelian. Ulasan singkat tersebut

dikategorikan memiliki tingkat *low concreteness* (kekonkretan rendah), yang menurut Azimi dan Andonova [2] merupakan konten abstrak yang hanya mengungkapkan perasaan pribadi tanpa memberikan detail mengenai produk. Hal ini didukung oleh temuan Bilal dan Almazroi [3] yang menyatakan bahwa panjang ulasan merupakan fitur tekstual yang sangat penting, di mana ulasan yang tidak bermanfaat secara signifikan memiliki panjang rata-rata yang lebih pendek dibandingkan ulasan yang bermanfaat. Kriteria ini ditetapkan secara *rule-based* sebelum tahap pembersihan teks guna menjaga keaslian struktur sintaksis ulasan mentah.

Hal ini diperburuk oleh pengguna yang cenderung mengejar *reward* berupa poin dengan memberikan ulasan singkat seperti 'Sip', 'Ok', atau sekadar karakter acak yang tidak sesuai dengan kualitas produk. Kecenderungan penggunaan kata yang tidak konkret ini membuat informasi dalam ulasan menjadi kurang diagnostik bagi pembaca. Sejalan dengan penelitian Alamsyah dkk. [4] yang menyoroti adanya kecenderungan ulasan tidak relevan atau palsu pada platform Tokopedia yang sulit dipahami persepsi dari konsumennya. Ulasan tidak informatif menimbulkan dampak berkelanjutan secara langsung pada keseluruhan platform *e-commerce*. Bagi calon pembeli, keberadaan ulasan yang hanya berisi simbol, kata singkat, atau informasi yang tidak sesuai dengan produk dapat menimbulkan kebingungan dalam menilai kualitas produk [4]. Sebagaimana ditekankan dalam penelitian Bilal dan Almazroi [3], tumpukan ulasan berkualitas rendah tersebut mengakibatkan masalah kelebihan informasi (*information overload*) yang menghambat kemampuan pelanggan dalam mengevaluasi kualitas produk saat membuat keputusan. Di sisi lain, penjual yang jujur dapat dirugikan karena ulasan informatif mereka tertutup oleh ulasan tidak bermakna. Selain itu, bagi pihak Tokopedia, ulasan tidak informatif menyebabkan penurunan kualitas data, sehingga mengganggu proses evaluasi produk, analisis perilaku konsumen, dan pengembangan sistem rekomendasi berbasis data [5].

Pendekatan *Knowledge Discovery in Database* (KDD) telah diterapkan dalam berbagai permasalahan *e-commerce* untuk mengekstraksi pola tersembunyi dari data kompleks dan menghasilkan pengetahuan bisnis yang dapat ditindaklanjuti [6]. Penerapan kerangka kerja KDD dalam penelitian ini menjadi krusial karena deteksi ulasan tidak informatif bukan hanya sekedar klasifikasi biner, melainkan proses penemuan pola (p*attern discovery*) pada data tekstual yang tidak terstruktur. Melalui tahapan KDD yang terstruktur, integrasi *K-Means* dan *Random Forest* tidak hanya berfungsi sebagai model prediktif, tetapi juga sebagai instrumen untuk mengidentifikasi karakteristik ulasan yang tidak memberikan nilai diagnostik bagi pengguna platform. Kerangka kerja ini relevan untuk diterapkan dalam proses klasifikasi ulasan guna meningkatkan kualitas informasi yang dihasilkan. Selain itu, integrasi metode interpretasi seperti SHAP (*SHapley Additive exPlanations*) sangat krusial dalam memastikan transparansi keputusan model, sehingga setiap klasifikasi ulasan tidak hanya akurat secara statistik tetapi juga dapat dipahami secara manajerial.

Pemanfaatan teknologi *Machine Learning* dan *Text Mining* menjadi solusi krusial dalam mengatasi tantangan analisis data besar pada platform Tokopedia, di mana volume ulasan harian yang masif tidak memungkinkan untuk diproses secara manual. Penelitian ini mengimplementasikan algoritma *K-Means*, sebuah metode *unsupervised learning* yang mengelompokkan data berdasarkan fitur tekstual secara iteratif [7]. Implementasi *clustering* ini bertujuan untuk mengidentifikasi pola tersembunyi serta menyaring *outlier* agar model dapat berfokus pada data yang memiliki karakteristik kelompok yang jelas [8]. Selanjutnya hasil pengelompokan tersebut diintegrasikan dengan algoritma *Random Forest* untuk mengklasifikasikan ulasan tidak informatif secara otomatis. Terkait hal tersebut, Inonu dkk. [9] menjelaskan bahwa keunggulan *Random Forest* meliputi kemampuan menangani data dengan dimensi tinggi, toleransi terhadap *overfitting*, serta pengurangan varians dalam prediksi. Secara teknis, keunggulan ini didasarkan pada mekanisme kerja algoritma yang melakukan *bagging* dan pemilihan fitur secara acak. Melalui proses tersebut, risiko *overfitting* dapat diminimalisir sehingga akurasi klasifikasi meningkat dibandingkan penggunaan *decision tree* tunggal. Dalam konteks pemilihan algoritma, penggunaan *Random Forest* menjadi sangat relevan karena keunggulannya sebagai metode *ensemble learning* yang memiliki ketahanan (*robustness*) tinggi terhadap *noise* pada data tekstual. Penelitian terbaru oleh Fahria dkk. [10] mengonfirmasi bahwa *Random Forest* memberikan performa terbaik dengan nilai akurasi mencapai 0,84 dibandingkan algoritma *boosting* lainnya dalam menangkap pola perilaku digital. Keunggulan tersebut didasarkan pada mekanisme kombinasi pohon keputusan yang secara efektif mereduksi varians dan meningkatkan stabilitas prediksi pada *dataset* yang besar. Dengan pendekatan KDD, integrasi teknologi ini diharapkan mampu memberikan performa deteksi yang tangguh dalam memisahkan ulasan bermakna dari gangguan informasi yang tidak relevan.

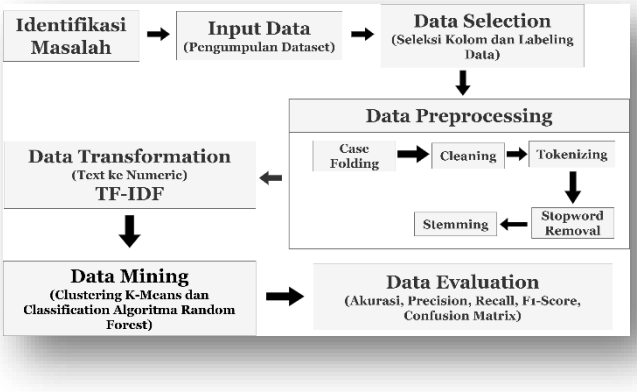
Penelitian sebelumnya yang telah dilakukan oleh Angkoso, C. V dkk. [8] menunjukan bahwa performa klasifikasi dapat ditingkatkan secara signifikan melalui integrasi teknik *K-Means Clustering* sebagai tahap pra-pemrosesan. Dalam studinya, model *Random Forest* tanpa optimasi hanya menghasilkan akurasi sebesar 77,49%, namun setelah diintegrasikan dengan *clustering* untuk menyaring *outlier* dan mengidentifikasi pola tersembunyi, akurasi melonjak drastis hingga mencapai 98,47%. Hal ini didukung oleh penelitian Al-Abadi dkk. [11] meskipun diterapkan pada domain yang berbeda yaitu keamanan siber, kerangka kerja ERF-KMC (*Enhanced Random Forest Classifier with K-Means Clustering*). Penelitian tersebut membuktikan bahwa penggunaan *K-Means* untuk membagi data ke dalam beberapa *cluster* terbukti secara signifikan meningkatkan akurasi dan stabilitas model dalam menangani data dalam jumlah besar. Proses pengelompokan ini memungkinkan algoritma untuk

bekerja pada segmen data yang lebih homogen dan terstruktur, sehingga model dapat lebih fokus dalam mengenali pola-pola unik pada setiap kelompok. Integrasi kedua metode ini terbukti memberikan kontribusi krusial terhadap peningkatan nilai akurasi prediktif dan generalisasi model klasifikasi. Hal ini diperkuat oleh hasil penelitian Priyanto dkk. [12] yang membuktikan bahwa optimasi *Random Forest* menggunakan *K-Means* secara signifikan meningkatkan performa model hingga mencapai akurasi 98,41% dengan cara mengatasi masalah ketidakseimbangan data dan menyederhanakan struktur data. Melengkapi temuan tersebut, Fiddin dkk. [13] juga telah menerapkan metode *supervised learning* untuk memvalidasi ulasan pada aplikasi Tokopedia dengan tingkat akurasi sebesar 77,42%.

Meskipun penelitian terdahulu menunjukkan potensi besar dalam analisis teks, fokus spesifik pada deteksi ulasan tidak informatif umumnya masih bertumpu pada optimasi algoritma menggunakan fitur leksikal konvensional. Pendekatan ini memiliki keterbatasan karena data tekstual (*natural language*) pada dasarnya bersifat *sparse* dan memiliki dimensionalitas yang sangat tinggi, sehingga model seringkali kehilangan konteks makna dari ulasan. Oleh karena itu, penelitian ini tidak menggunakan *Davies-Bouldin Index* (DBI) sebagai tolak ukur kualitas klaster secara absolut, melainkan sebagai metrik komparatif relatif. Celah dalam penelitian ini terletak pada integrasi *K-Means* dan *Random Forest* melalui kerangka KDD, di mana *K-Means* difungsikan secara spesifik sebagai instrumen penemuan pengetahuan (*knowledge discovery*) untuk memetakan secara alami pola topik ulasan konsumen. Melalui penggabungan fitur leksikal dan fitur topik laten ini, algoritma *Random Forest* memiliki representasi data yang lebih kaya. Penelitian ini akan membandingkan performa model *Random Forest* murni dengan model hibrida *Random Forest* dan *K-Means*, yang diharapkan dapat menghasilkan pengetahuan baru mengenai peningkatan performa deteksi ulasan tidak informatif secara lebih akurat.

II. METODE

Data yang ada di ulasan pada platform Tokopedia muncul dalam format teks yang tidak terstruktur, sehingga dibutuhkan beberapa langkah pengolahan untuk informasi yang penting dalam mengidentifikasi ulasan yang tidak memberikan informasi. Tahapan pengolahan data ini dilakukan dengan cara yang teratur berdasarkan pendekatan KDD, yang dioptimasi dengan integrasi *K-Means Clustering* sebagai tahap pra-pemrosesan. Tahapan penelitian meliputi seleksi, pra-pemrosesan, dan transformasi data, di mana algoritma *K-Means* berfungsi mereduksi *outlier* untuk meningkatkan kualitas *dataset* pelatihan. Kualitas klaster kemudian dievaluasi menggunakan *Davies-Bouldin Index* (DBI) sebelum diklasifikasikan melalui algoritma *Random Forest*.



Gambar 1. Diagram Alur Proses Klasifikasi

Performa model diukur secara komprehensif pada tahap evaluasi akhir, dengan detail alur proses yang disajikan pada Gambar 1.

A. Teknik Pengumpulan Data

Teknik pengumpulan data yang digunakan dalam penelitian ini berasal dari kumpulan *dataset* yang terdapat di situs Kaggle dengan tautan: https://www.kaggle.com/datasets/salmanabdu/tokopedia-product-reviews-2025?utm. *Dataset* ini didapatkan melalui penyalinan tabel data dan kemudian disimpan sebagai tabel *dataset* yang mencakup sekitar 65.543 ulasan pelanggan Tokopedia *Product Reviews* 2025 dalam bentuk teks beserta metadata terkait. *Dataset* ini akan diproses lebih lanjut untuk klasifikasi ulasan tidak informatif, rincian deskripsi *dataset* disajikan pada Tabel I.

TABEL I
DESKRIPSI DATASET

Nama Kolom	Deskripsi
review_text	Konten mentah dari ulasan pengguna, termasuk emoji, simbol, dan tanda baca asli.
review_date	Tanggal pengiriman ulasan oleh pengguna (format: YYYY-MM-DD).
review_id	Identitas unik alfanumerik untuk setiap ulasan guna mencegah duplikasi data.
product_name	Judul lengkap dari produk yang tertera pada daftar penjualan.
product_category	Kategori utama produk (contoh: Elektronik, Kesehatan, Fashion).
product_variant	Detail spesifik varian produk yang dibeli (seperti warna, ukuran, atau model).
product_price	Harga jual produk dalam satuan mata uang Rupiah (IDR).
product_url	Alamat tautan (URL) langsung menuju halaman produk di platform Tokopedia.
product_id	Identitas unik yang ditetapkan untuk daftar produk tertentu.
rating	Skor penilaian numerik yang diberikan oleh pelanggan dengan skala 1 sampai 5.
sold_count	Total jumlah unit produk yang telah terjual pada daftar tersebut.

shop_id	Identitas unik toko/penjual yang telah dianonimkan untuk menjaga privasi merchant.
sentiment_label	Klasifikasi sentimen dari ulasan yang dibagi menjadi Positif, Netral, atau Negatif.

B. Tahapan Knowledge Discovery in Database

Penelitian ini mengikuti prosedur KDD secara berurutan untuk menghasilkan olahan data yang valid dan dapat dipertanggungjawabkan. Terdapat lima langkah utama dalam proses KDD yaitu *Data Selection, Preprocessing, Transformation, Data Mining*, dan *Evaluation*. Rincian dari alur kerja tersebut dijelaskan sebagai berikut.

1) Data Selection: Tahap seleksi data dilakukan dengan menyaring *dataset* yang digunakan guna memperoleh variabel yang krusial. Atribut-atribut yang dipilih merupakan parameter penting yang secara teknis dianggap paling menentukan dalam prediksi [14]. Bersamaan dengan *Data Selection*, pelabelan awal (*ground truth*) dilakukan pada ulasan mentah menggunakan pendekatan berbasis aturan (*rule-based*) [15]. Tahap ini diproses sebelum pembersihan teks guna menjaga keaslian struktur sintaksis ulasan mentah, seperti jumlah kata.

2) Data Preprocessing: Preprocessing merupakan tahapan krusial dalam pengolahan data teks tidak terstruktur yang berfungsi untuk menyeleksi data pada setiap dokumen yang akan diproses [16]. Tahapan ini bertujuan untuk menghilangkan gangguan (*noise*) seperti *emoticon*, angka, karakter khusus, dan kata-kata yang tidak diperlukan guna meningkatkan kualitas data serta akurasi klasifikasi [17], [18]. Proses ini menyederhanakan teks melalui serangkaian langkah sistematis yang umumnya meliputi *cleaning, case folding, tokenizing, stopword removal* dan *stemming* untuk mereduksi kata umum serta mengembalikan kata ke bentuk dasarnya [1], [17], [19].

3) Data Transformation: Transformasi data merupakan langkah pengubahan format data agar sesuai dengan prasyarat algoritma klasifikasi yang digunakan. Tahapan ini bertujuan untuk menyelaraskan karakteristik setiap atribut data dengan kebutuhan teknis model, sehingga proses penambangan data (*data mining*) dapat berjalan lebih efektif dan optimal. Dalam algoritma *Random Forest*, ulasan teks perlu diubah menjadi angka atau fitur baru yang lebih representatif agar model dapat mengenali pola informasi dengan lebih akurat [14], [15].

4) Data Mining: Data Mining adalah adalah proses untuk menemukan informasi berharga dari kumpulan data yang besar, baik yang terstruktur maupun yang tidak terstruktur, dengan memanfaatkan metode atau algoritma tertentu. Pada fase ini, dilakukan pengembangan model klasifikasi melalui penggunaan algoritma *Random Forest* yang dioptimalkan dengan tahap *preprocessing* menggunakan *K-Means Clustering*. *K-Means Clustering* diterapkan untuk mengelompokkan data ke dalam klaster-klaster berdasarkan kesamaan karakteristiknya. Pendekatan ini bertujuan untuk menyaring *outlier* dan memastikan bahwa hanya kelompok data dengan pola yang jelas yang digunakan, sehingga menghasilkan data *training* yang jauh lebih berkualitas [8]. Setelah data dikelompokkan dan dibersihkan, proses pengklasifikasian akhir dilakukan menggunakan algoritma *Random Forest. Random Forest* adalah teknik dalam *machine learning* yang menggabungkan beberapa *Decision Tree* untuk mengelompokkan data ke dalam kategori tertentu. Algoritma ini dipilih karena kehandalannya dalam mengelola volume data yang besar, menangani banyak variabel, serta kemampuannya untuk meminimalisir risiko *overfitting* dari berbagai pohon keputusan yang terbentuk [15].

5) Evaluation: Tahap *evaluation* berfungsi untuk menyaring hasil dari *data mining* menjadi informasi yang aplikatif dengan menyajikan pola yang ditemukan dalam format yang jauh lebih mudah dipahami, sehingga kesimpulan penelitian jadi lebih jelas [15]. Evaluasi utama difokuskan pada penentuan tingkat performa klasifikasi menggunakan matrix *Accuracy*, *Confusion Matrix*, *precision*, *recall*, dan *f1-score* [19], dengan ketentuan penilaian merujuk pada format Tabel II.

TABEL II
CONFUSION MATRIX

Predicted	***Actual***	
	True	***False***
True	TP (*True Positive)*	FP *(False Positive)*
False	TN *(True Negative)*	FN *(False Negative)*

Pengujian stabilitas model dilakukan menggunakan skema *Train-Test Split* yang divalidasi lebih lanjut dengan *5-Fold Cross-Validation.* Selanjutnya, performa model *hybrid* dikomparasikan secara langsung dengan model *baseline* (*Random Forest* standar). Serta mengimplementasikan metode SHAP (*SHapley Additive exPlanations*) sebagai teknik interpretabilitas (*feature importance*) guna membedah dan memahami bagaimana setiap fitur mempengaruhi keputusan model. Kemudian ditutup dengan analisis kesalahan (*error analysis*) secara mendalam untuk mengidentifikasi pola ulasan yang gagal diklasifikasikan, sehingga batasan sistem dapat dipahami secara transparan.

III. HASIL DAN PEMBAHASAN

Bagian ini menguraikan hasil dari implementasi kerangka kerja KDD untuk mendeteksi ulasan tidak informatif. Pembahasan disusun secara berurutan mengikuti tahapan KDD, mulai dari tahap seleksi data, pra-pemrosesan teks, hingga transformasi fitur numerik menggunakan metode TF-IDF. Pada tahap inti (*Data Mining*), performa model

klasifikasi *hybrid* yang menggabungkan algoritma *K-Means* dan *Random Forest* dievaluasi secara komprehensif. Pengujian tidak hanya bersandar pada metrik performa standar, tetapi juga mencakup perbandingan *baseline*, uji stabilitas (*5-Fold Cross-Validation*), analisis kesalahan (*error analysis*), serta interpretabilitas model menggunakan metode SHAP.

A. Data Selection

Pengumpulan dan penyeleksian *dataset* "Tokopedia Product Reviews 2025" merupakan tahap pertama dari penelitian ini. *Dataset* disederhanakan dengan mempertahankan atribut kolom teks ulasan untuk analisis. Agar efisiensi komputasi pada pemodelan *Machine Learning* tetap terjaga tanpa menghilangkan representasi data, dilakukan *random sampling* sebanyak 35.000 data, dilanjutkan dengan pembersihan baris data kosong (*missing values*).

B. Data Preprocessing

Tahap pra-pemrosesan data bertujuan untuk membersihkan teks ulasan mentah agar lebih terstruktur dan siap diproses oleh algoritma *Machine Learning*. Berdasarkan sistem yang dibangun, proses ini dibagi menjadi enam tahapan berurutan.

1) *Labelling:* Tahapan *labeling* ini dilakukan untuk memberikan target kelas (1 untuk informatif dan 0 untuk tidak informatif) pada setiap teks berdasarkan aturan karakteristik teksnya (*rule-based*). Pelabelan awal (*ground truth)* dilakukan pada ulasan mentah menggunakan pendekatan berbasis aturan sebelum pembersihan teks guna menjaga keaslian struktur sintaksis, seperti jumlah kata asli. Kriteria pelabelan ini merujuk pada temuan Bilal dan Almazroi [3] yang menyatakan bahwa panjang ulasan merupakan fitur tekstual yang krusial, di mana ulasan tidak bermanfaat memiliki panjang rata-rata yang lebih pendek. Selain itu, ulasan yang terdeteksi hanya berisi pernyataan subjektif tanpa detail atribut produk dikategorikan sebagai *low concreteness* (kekonkretan rendah) merujuk pada teori Azimi dan Andonova [2]. Untuk menjamin kualitas label yang dihasilkan, proses validasi dilakukan secara manual oleh dua annotator independen terhadap sampel representatif sebanyak 100 ulasan yang dipilih secara acak. Konsistensi pelabelan antar annotator diukur menggunakan *inter-annotator agreement* dengan nilai *Cohen's Kappa* sebesar 0,947, yang menunjukkan tingkat kesepakatan yang sangat kuat. Hal ini memastikan bahwa ulasan yang dikategorikan tidak informatif bersifat tidak diagnostik bagi konsumen. Hasil pelabelan disajikan pada Tabel III.

TABEL III
HASIL *LABELLING*

No	Ulasan Mentah	Label
1	Sayur & lauk semua fresh, respon cepat dan...	1
2	Bagus..langsung pakai, mantabbb	0

2) *Case Folding:* Tahapan *case folding* bertujuan mengubah Keseluruhan teks menjadi huruf kecil (*lowercase*) untuk mencegah sistem membedakan makna suatu kata berdasarkan penggunaan huruf kapital saat dianalisis. Hasil dari proses ini disajikan pada Tabel IV.

TABEL IV
HASIL *CASE FOLDING*

No	Sebelum Proses	Setelah Proses
1	Sayur & lauk semua fresh, respon cepat dan...	sayur & lauk semua fresh, respon cepat dan...
2	Bagus..langsung pakai, mantabbb	bagus..langsung pakai, mantabbb

3) *Cleaning:* Tahap ini berfungsi untuk menghapus karakter-karakter yang tidak memiliki nilai semantik atau informasi penting, seperti tautan (URL), angka, dan seluruh tanda baca. Pada tahap ini, simbol khusus atau entitas HTML juga disesuaikan. Hasil dari tahapan ini disajikan pada Tabel V.

TABEL V
HASIL *CLEANING*

No	Sebelum Proses	Setelah Proses
1	sayur & lauk semua fresh, respon cepat dan...	sayur lauk semua fresh respon cepat dan
2	bagus..langsung pakai, mantabbb	bagus langsung pakai mantabbb

4) *Tokenizing: Tokenizing* adalah proses memecah kalimat menjadi kata-kata tunggal (*token*) agar dapat dihitung dan dianalisis secara individual oleh model. Data hasil pemecahan kalimat ini tertera pada Tabel VI.

TABEL VI
HASIL *TOKENIZING*

No	Sebelum Proses	Setelah Proses
1	sayur lauk semua fresh respon cepat dan	['sayur', 'lauk', 'semua', 'fresh', 'respon', 'cepat', 'dan']
2	bagus langsung pakai mantabbb	['bagus', 'langsung', 'pakai', 'mantabbb']

5) *Stopword Removal:* Tahap ini bertujuan untuk membuang kata-kata hubung atau kata umum yang sering muncul tetapi tidak memberikan makna spesifik terhadap ulasan (*stopword*), seperti "dan", "di", "yang". Pada penelitian ini, digunakan *library* Sastrawi untuk menyaring kata-kata tersebut guna mengurangi dimensi data. Perbandingan data sebelum dan sesudah proses dimuat pada Tabel VII.

TABEL VII
HASIL *STOPWORD REMOVAL*

No	Sebelum Proses	Setelah Proses
1	['sayur', 'lauk', 'semua', 'fresh', 'respon', 'cepat', 'dan']	['sayur', 'lauk', 'semua', 'fresh', 'respon', 'cepat']
2	['bagus', 'langsung', 'pakai', 'mantabbb']	['bagus', 'langsung', 'pakai', 'mantabbb']

6) *Stemming:* Tahap *stemming* dilakukan untuk mengembalikan kata yang memiliki imbuhan menjadi bentuk kata dasarnya. Hal ini mengelompokkan kata-kata yang memiliki akar makna yang sama agar mesin lebih fokus pada inti kata. Hasil pengembalian kata dasar ini dapat dilihat pada Tabel VIII.

TABEL VIII
HASIL *STEMMING*

No	Sebelum Proses	Setelah Proses
1	['sayur', 'lauk', 'semua', 'fresh', 'respon', 'cepat']	['sayur', 'lauk', 'semua', 'fresh', 'respon', 'cepat']
2	['bagus', 'langsung', 'pakai', 'mantabbb']	['bagus', 'langsung', 'pakai', 'mantab']

C. Data Transformation

Setelah data diolah secara menyeluruh pada tahap pra-pemrosesan, langkah berikutnya adalah mengkonversi data teks bersih tersebut menjadi bentuk numerik agar algoritma klasifikasi dapat memprosesnya. Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) diterapkan untuk mengubah representasi teks menjadi matriks angka yang menggambarkan seberapa penting sebuah kata di dalam keseluruhan dokumen ulasan [6]. Proses vektorisasi ini dieksekusi dengan konfigurasi *n-gram* pada tingkat *unigram* (1,1) guna menangkap bobot spesifik dari setiap kata tunggal. Dalam memastikan efisiensi komputasi dan mereduksi dimensi data dari *noise* teks yang tidak relevan, ekstraksi dibatasi secara ketat pada 1.000 fitur (*max_features*). Transformasi ini berhasil memadatkan informasi krusial dari korpus teks menjadi matriks berdimensi 35.000 baris ulasan dan 1.000 kolom fitur. Sepuluh kata teratas dengan total bobot TF-IDF tertinggi dari hasil ekstraksi tersebut disajikan pada Tabel IX.

TABEL IX
HASIL EKSTRAKSI FITUR TF-IDF

No	Kata	Frekuensi	Nilai TF-IDF
1	sesuai	8621	2075.485116
2	barang	8367	1938.158535
3	bagus	7565	1931.943791
4	cepat	8775	1917.409263
5	pengiriman	6157	1451.504103

Nilai yang terdapat dalam Tabel IX memperlihatkan kata-kata dengan total bobot tertinggi yang berkontribusi besar pada pembuatan model klasifikasi. Ini menggarisbawahi bahwa ulasan dari konsumen sangat menekankan pada kesesuaian serta mutu produk (dilihat dari kata 'sesuai', 'barang', dan 'bagus'), dan juga kecepatan layanan pengiriman barang (terwakili oleh kata 'cepat' dan 'pengiriman'). Faktor-faktor spesifik inilah yang menjadi petunjuk utama bagi algoritma untuk menentukan apakah sebuah ulasan bersifat informatif atau tidak informatif.

D. Data Mining

Pada tahap *Data Mining*, proses pemodelan dilakukan dengan menggabungkan dua algoritma, yaitu *K-Means* untuk tahap *clustering* (pengelompokan) dan *Random Forest* untuk tahap klasifikasi akhir. Langkah pertama adalah melakukan *clustering* menggunakan algoritma *K-Means* terhadap representasi teks yang telah diekstraksi melalui TF-IDF (berjumlah 1.000 fitur dari 35.000 baris data teks bersih). Untuk menentukan jumlah klaster (*k*) yang paling optimal, penelitian ini mengevaluasi kualitas klaster menggunakan *Davies-Bouldin Index* (DBI). Pengujian dilakukan secara iteratif dengan nilai *k* mulai dari 2 hingga 8. Semakin rendah skor DBI, semakin baik pemisahan antar klaster yang dihasilkan. Hasil evaluasi penentuan jumlah klaster disajikan pada Tabel X.

TABEL X
HASIL EVALUASI DAVIES-BOULDIN INDEX (DBI)

Jumlah *Cluster* (k)	Skor DBI
2	5,6267
3	4,9962
4	4,6826
5	4,3879
6	**4,2402**
7	4,2615
8	4,4318

Berdasarkan Tabel X, jumlah *cluster* optimal ditentukan pada $k = 6$ dengan skor DBI terendah sebesar 4,2402. Perlu dicatat bahwa secara absolut, nilai DBI ini tergolong relatif tinggi, yang mengindikasikan adanya tumpang tindih antar-klaster pada ruang fitur berdimensi tinggi. Kondisi ini merupakan karakteristik lumrah dari pemrosesan data teks (*natural language*) yang memiliki sifat *sparse* (jarang) dan dimensionalitas yang sangat tinggi, sehingga pemisahan klaster yang sempurna secara geometris sangat sulit dicapai. Oleh karena itu, dalam penelitian ini, kami tidak menggunakan DBI sebagai tolak ukur kualitas *cluster* secara absolut, melainkan sebagai metrik komparatif relatif. *K-Means* difungsikan bukan sekadar untuk segmentasi spasial, melainkan sebagai instrumen penemuan pengetahuan (*knowledge discovery*) untuk memetakan secara alami pola topik ulasan konsumen seperti klaster yang spesifik membahas kualitas produk berbanding efisiensi logistik. Hasil penempatan setiap teks ulasan ke dalam 6 klaster tersebut kemudian direpresentasikan secara eksplisit sebagai label *cluster* diskret (dengan nilai 0 hingga 5). Label ini kemudian ditambahkan sebagai satu kolom fitur baru (fitur prediktor) ke dalam matriks TF-IDF awal. Representasi berupa label *cluster* dipilih karena ia secara langsung mengkodekan keanggotaan kelompok tekstual setiap ulasan, memberikan konteks struktural mengenai pola kemiripan antar-teks kepada algoritma pengklasifikasi. Penggabungan ini menghasilkan matriks *hybrid* berdimensi 1.001 (terdiri dari 1.000 fitur ekstraksi kata TF-IDF dan 1 fitur *label cluster K-Means*), yang selanjutnya digunakan sebagai *input* pelatihan model algoritma *Random Forest* dengan jumlah data latih sebanyak 27.984 sampel (80% dari *dataset*).

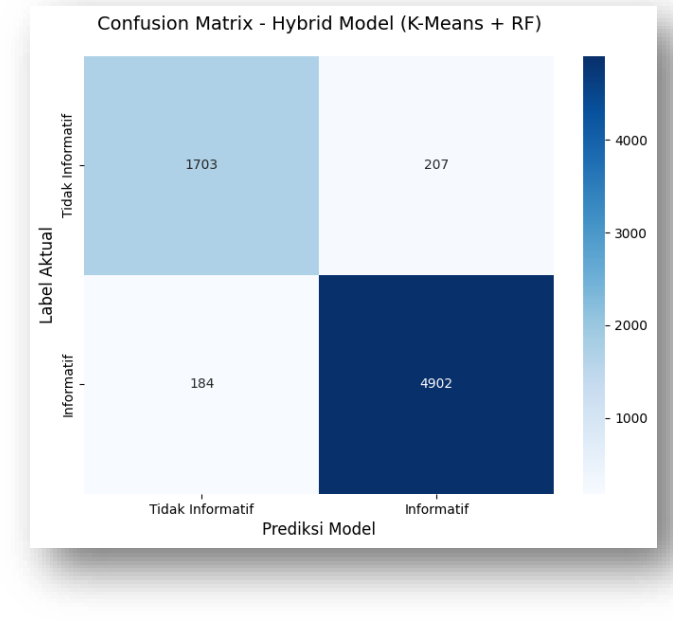
E. Data Evaluation

Bagian ini menyajikan evaluasi komprehensif terhadap kinerja model *hybrid* yang diusulkan dalam mengklasifikasikan ulasan *e-commerce*. Pengujian dilakukan menggunakan data uji sebesar 20% dari total *dataset*, yang setara dengan 6.996 sampel ulasan independen. Matrik evaluasi yang digunakan mencakup *Accuracy*, *Precision*, *Recall*, dan *F1-score* untuk memberikan gambaran performa model secara menyeluruh dari berbagai sudut pandang statistik. Hasil evaluasi klasifikasi disajikan secara terperinci pada Tabel XI.

TABEL XI
LAPORAN KLASIFIKASI MODEL

Kelas Target	*precision*	*recall*	*F1-Score*	*Support*
Tidak Informatif (0)	0,90	0.89	0.90	1910
Informatif (1)	0.96	0.96	0.96	5086
Akurasi Keseluruhan			**0.94 (94,4%)**	

Berdasarkan Tabel XI, model *hybrid* menunjukkan performa yang sangat stabil di kedua kelas. Hal yang paling signifikan untuk disoroti adalah kemampuan model dalam mempertahankan nilai *F1-score* sebesar 0,90 pada kelas minoritas ("Tidak Informatif"). Hasil pengujian membuktikan ketahanan (*robustness*) model terhadap kondisi ketidakseimbangan data (*imbalanced data*). Meskipun jumlah data ulasan informatif jauh lebih banyak, model mampu mengenali ulasan tidak informatif dengan tingkat presisi dan sensitivitas yang proporsional.



Gambar 2. *Confusion Matrix*

Distribusi hasil klasifikasi antara label aktual dan label prediksi untuk setiap kelas disajikan pada Gambar 2. Visualisasi pada Gambar 2 mengonfirmasi bahwa tingkat kesalahan klasifikasi model sangat rendah. Model berhasil memprediksi dengan tepat 1.703 ulasan sebagai kategori "Tidak Informatif" dan 4.902 ulasan sebagai kategori "Informatif". Margin kesalahan yang sangat kecil, dimana hanya terdapat 207 kasus *False Positive* dan 184 kasus *False Negative* dari total hampir tujuh ribu data uji. Integrasi fitur klaster *K-Means* ke dalam arsitektur *Random Forest* menciptakan batas keputusan (*decision boundary*) yang sangat jelas dan akurat dalam membedakan karakteristik teks ulasan. Adapun pengujian 5-*Fold Cross-Validation* disajikan pada Tabel XII.

TABEL XII
HASIL PENGUJIAN 5-FOLD CROSS-VALIDATION

Iterasi Pengujian	Tingkat Akurasi (%)
Fold 1	94.63
Fold 2	94.7
Fold 3	94.57
Fold 4	94.71
Fold 5	94.18
Rata-rata Keseluruhan	**94.56**
Standar Deviasi	**±0,19**

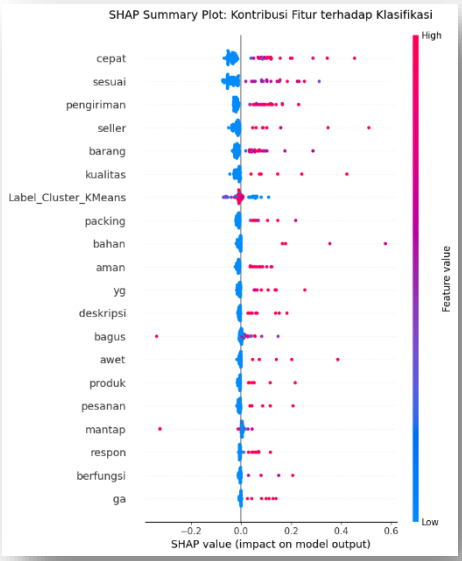
Berdasarkan Tabel XII, pengujian *5-Fold Cross-Validation* diimplementasikan untuk memberikan validasi tambahan bahwa tingkat akurasi yang dilaporkan sebelumnya bukanlah hasil dari anomali pembagian data (*data leakage* atau bias partisi). Metode ini secara komprehensif menguji keandalan arsitektur dengan mengevaluasi model pada lima subset (lipatan) data yang berbeda dan independen secara bergantian. Hasil evaluasi menunjukkan bahwa model *hybrid* secara konsisten mampu mempertahankan performa tertingginya pada seluruh skenario iterasi. Model mencatatkan rata-rata akurasi sebesar 94,56% dengan tingkat variansi yang sangat minim. Angka standar deviasi yang tergolong sangat kecil (±0,19%) tersebut memberikan bukti empiris yang solid terkait stabilitas algoritma. Pencapaian ini secara konklusif menjawab batasan evaluasi model *machine learning*. Arsitektur yang diusulkan terbukti memiliki kemampuan generalisasi yang andal, kebal terhadap indikasi *overfitting*, serta mampu beroperasi secara konsisten pada berbagai variasi distribusi data ulasan *e-commerce*.

Penelitian ini juga melakukan analisis komparatif untuk menjustifikasi signifikansi integrasi *K-Means* dalam kerangka kerja KDD, antara model *hybrid* dengan algoritma *baseline* yang lebih sederhana, yakni *Random Forest* murni. Hasil perbandingan kinerja kedua model disajikan pada Tabel XIII.

TABEL XIII
PERBANDINGAN KINERJA MODEL

Model	Akurasi
Baseline	94,31%
Hybrid	94,41%

Berdasarkan Tabel XIII, analisis komparatif dilakukan untuk melihat nilai tambah dari kerangka kerja *hybrid* dibandingkan algoritma *baseline* yang lebih sederhana. Model *baseline* telah menunjukkan performa awal yang sangat tinggi (94,31%), mengindikasikan bahwa ekstraksi fitur pada teks tersebut telah mendekati batas efektivitas maksimalnya (*ceiling effect*). Meskipun integrasi fitur klaster *K-Means* hanya memberikan peningkatan *marginal* sebesar 0,10%, fokus utama dari kerangka kerja ini bukanlah semata-mata mengejar metrik akurasi. Tujuan fundamental dari injeksi algoritma *K-Means* adalah sebagai instrumen eksplorasi (*knowledge discovery*). Hasil komparasi ini menjadi bukti empiris bahwa penambahan fitur klaster memberikan dimensi interpretasi baru pada model tanpa mendegradasi atau merusak performa tinggi dari *baseline*. Atribut klaster yang dihasilkan sukses menangkap pola semantik laten ulasan yang tidak terjangkau oleh fitur TF-IDF konvensional, sehingga model *hybrid* tidak hanya akurat, tetapi juga memiliki tingkat interpretabilitas yang lebih kaya.



Gambar 3. SHAP (*Shapley Additive exPlanations)*

Penerapan metode SHAP (*SHapley Additive exPlanations*) bertujuan untuk menginterpretasikan proses pengambilan keputusan algoritma *black-box* secara visual, yang hasilnya disajikan pada Gambar 3. Berdasarkan Gambar 3, hasil ekstraksi *feature importance* menampilkan 20 fitur dengan kontribusi terbesar terhadap hasil prediksi. Kosakata yang berkaitan dengan logistik dan kualitas produk, seperti 'cepat', 'sesuai', dan 'pengiriman', secara konsisten menempati posisi teratas. Temuan yang paling menonjol adalah fitur *Label_Cluster_KMeans* yang berhasil menduduki peringkat ke-7 dari keseluruhan 1.001 fitur prediktor. Hal ini membuktikan bahwa representasi label klaster memberikan informasi topik struktural yang krusial bagi algoritma *Random Forest* dalam membedakan ulasan informatif dan tidak informatif

Meskipun model *hybrid* yang diusulkan mencapai tingkat akurasi yang sangat tinggi, evaluasi kerangka KDD menuntut adanya eksplorasi terhadap batasan sistem. Oleh karena itu, dilakukan analisis kesalahan (*error analysis*) terhadap 391 sampel ulasan yang gagal diklasifikasikan dengan benar, yakni 207 kasus *False Positive* (FP) dan 184 kasus *False Negative* (FN). Analisis kualitatif terhadap sampel kesalahan tersebut disajikan pada Tabel XIV.

TABEL XIV
EKSTRAKSI DAN ANALISIS KESALAHAN PREDIKSI MODEL

Kategori	Contoh Teks Ulasan	Analisis Penyebab Kesalahan
False Positive (FP)	"produk oke, semoga semangat olahraganya sama…"	Model terdistraksi oleh panjang kalimat, repetisi kata positif, dan gagal mendeteksi konteks sarkasme atau ulasan minim substansi produk.
False Negative (FN)	"bocor, tumpah, rusak"	Model menunjukkan bias terhadap ulasan sangat pendek atau penggunaan leksikon non-standar sehingga gagal menangkap konteks kerusakan.

Berdasarkan Tabel XIV, penemuan pengetahuan (*knowledge discovery*) tingkat lanjut berhasil mengungkap tantangan linguistik alami dalam ulasan berbahasa Indonesia. Pada kasus FP, algoritma klasifikasi sering kali terdistraksi oleh ulasan yang secara leksikal panjang atau mengandung sentimen positif (seperti "aman", "oke"), namun pada realitanya tidak mengandung informasi spesifik yang bermanfaat bagi calon pembeli lain. Sebaliknya, pada kasus FN, model menunjukkan bias terhadap ulasan yang sangat pendek. Kata-kata tunggal namun berbobot informasi tinggi terutama yang merujuk pada cacat fisik logistik seperti "bocor" atau "rusak" kerap diabaikan oleh model karena kurangnya konteks kalimat penyerta. Penemuan ini memberikan wawasan penting bahwa ambiguitas semantik, sarkasme, dan ulasan berkonteks rendah (*low-context*) masih menjadi batasan fundamental dalam pemrosesan bahasa alami. Analisis ini sekaligus membuktikan bahwa pendekatan evaluasi yang transparan sangat diperlukan untuk memahami karakteristik interaksi konsumen di platform *e-commerce* Indonesia secara utuh.

IV. KESIMPULAN

Penelitian ini berhasil mengimplementasikan kerangka kerja KDD melalui integrasi algoritma *K-Means Clustering* dan *Random Forest* untuk mendeteksi ulasan tidak informatif pada platform Tokopedia. Berdasarkan hasil eksperimen terhadap 35.000 data ulasan yang telah melalui proses

pembersihan, model *hybrid* yang diusulkan mencapai tingkat akurasi sebesar 94,41%. Kinerja model terbukti sangat stabil dengan nilai *F1-score* sebesar 0,96 untuk kelas informatif dan 0,90 untuk kelas tidak informatif. Pengujian stabilitas menggunakan *5-Fold Cross-Validation* memberikan validasi empiris yang kuat dengan rata-rata akurasi 94,56% dan standar deviasi yang sangat rendah yaitu ±0,19. Penemuan pengetahuan baru *(knowledge discovery)* dalam penelitian ini ditunjukkan melalui peran krusial fitur klaster yang dihasilkan oleh algoritma *K-Means*. Meskipun integrasi fitur klaster ini memberikan peningkatan akurasi marginal sebesar 0,10% dibandingkan model baseline (94,31%), fungsi utamanya adalah sebagai instrumen eksplorasi pola semantik. Analisis metode SHAP membuktikan signifikansi temuan ini, dimana fitur *Label_Cluster_KMeans* berhasil menempati peringkat ke-7 dari total 1.001 fitur prediktor dalam mempengaruhi keputusan model. Hal ini menunjukkan bahwa pengelompokan ulasan ke dalam 6 klaster berdasarkan pola topik laten (seperti kualitas produk versus efisiensi logistik) memberikan konteks struktural yang lebih kaya dibandingkan fitur TF-IDF konvensional. Selain itu, tahap penemuan pengetahuan tingkat lanjut melalui analisis kesalahan terhadap 207 kasus *False Positive* dan 184 kasus *False Negative* mengungkap bahwa adanya ambiguitas linguistik tetap menjadi tantangan utama. Model cenderung mengalami kesalahan pada ulasan yang panjang secara leksikal namun minim substansi fisik (sarkasme), serta mengabaikan ulasan sangat pendek yang memuat informasi kerusakan produk krusial seperti "bocor" atau "rusak". Secara manajerial, sistem ini memberikan kontribusi nyata bagi ekosistem Tokopedia dalam menjaga kualitas informasi dan mengatasi masalah kelebihan informasi *(information overload)* secara otomatis. Meskipun penelitian ini berfokus pada platform Tokopedia, kerangka kerja KDD yang diusulkan memiliki fleksibilitas tinggi untuk diadaptasi pada platform *e-commerce* lain maupun domain teks yang berbeda karena pendekatannya yang mengandalkan pola semantik laten yang bersifat universal. Pengembangan di masa depan diharapkan dapat menguji performa model pada dataset yang lebih bervariasi guna memperkuat validitas sistem dalam mendeteksi ulasan tidak informatif lintas platform secara lebih luas.

DAFTAR PUSTAKA

[1] W. Nurwahyudi, M. Isnaini, S. J. Roszi, dan A. W. Laily, "Analisis Sentimen Ulasan Produk Moisturizer Skintific Di Tokopedia Menggunakan Support Vector Machine," *Jurnal Sistem Informasi dan Bisnis Cerdas*, vol. 18, no. 1, hlm. 129–142, 2025.

[2] S. Azimi dan Y. Andonova, "Did you find this review helpful?," *Marketing Intelligence & Planning*, vol. 41, no. 3, hlm. 329-343, 2023.

[3] M. Bilal dan A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electronic Commerce Research*, vol. 23, no. 4, hlm. 2737-2757, 2023.

[4] H. Alamsyah, Y. Cahyana, dan A. R. Pratama, "Deteksi Fake Review Menggunakan Metode Support Vector Machine dan Naive Bayes Di Tokopedia," *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, vol. 12, no. 2, hlm. 585–598, 2023.

[5] P. Demetria dan A. Wedhasmara, "Analisis Sentimen Pelanggan Terhadap Penilaian Produk Pada Tokopedia Nyemil.Saji Menggunakan Metode Support Vector Machine," *JEMSI: Jurnal Ekonomi Manajemen Sistem Informasi*, vol. 7, no. 2, hlm. 1350–1361, 2025.

[6] D. A. Ardhani dan K. D. Tania, "Knowledge Discovery on E-Commerce Customer Churn Using Interpretable Machine Learning: A Comparative Study of SHAP-Based Classifiers," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, hlm. 2695–2702, Okt. 2025.

[7] N. Alfira, M. R. T. Ramdhani, M. R. P. Budika, M. V. Santoso, dan N. Zahry, "Analisis Sentimen Terhadap Komentar Negatif (Hate Speech) Di Twitter Dengan Algoritma K-Means Clustering Menggunakan RapidMiner," *Journal of Information Technology and Informatics Engineering*, vol. 1, no. 1, hlm. 57–61, 2025.

[8] C. V. Angkoso, M. A. N. Thrisna, B. D. Satoto, dan A. Kusumaningsih, "Optimasi Klasifikasi Sentimen Menggunakan Random Forest dengan Preprocessing K-Means Clustering dan SMOTE," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 10, no. 3, hlm. 389–400, 2024.

[9] O. Y. Inonu, K. Magda, dan A. Amarudin, "Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes," *Expert J. Manaj. Sist. Inf. dan Teknol.*, vol. 15, no. 1, hlm. 1, 2025.

[10] D. N. Fadhilahrizka, K. D. Tania, dan R. D. Kurnia, "Analisis Komparatif Algoritma Random Forest, XGBoost, dan CatBoost untuk Klasifikasi Tingkat Stres Pengguna Media Sosial," Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, vol. 11, no. 1, hlm. 1843–1853, 2026.

[11] A. A. J. Al-Abadi, M. B. Mohamed, dan A. Fakhfakh, "Enhanced Random Forest Classifier with K-Means Clustering (ERF-KMC) for Detecting and Preventing Distributed-Denial-of-Service and Man-in-the-Middle Attacks in Internet-of-Medical-Things Networks," *Computers*, vol. 12, no. 12, hlm. 262, Des. 2023.

[12] D. Priyanto, H. Hairani, K. Marzuki, dan M. Innuddin, "Optimization of Random Forest for Health Data Classification Using PCA and K-Means SMOTE-ENN," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, hlm. 27646–27652, 2025.

[13] F. Fiddin, M. Y. Syahbarna, dan M. Ridwan, "Penggunaan Supervised Learning untuk Prediksi Validitas Ulasan Negatif Aplikasi Tokopedia Berdasarkan Pengalaman Pengguna Ahli," *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, vol. 23, no. 2, hlm. 409–417, 2024.

[14] M. I. A. Rois, G. Dwilestari, dan N. Suarna, "Prediksi Persetujuan Pinjaman Menggunakan Dataset Loan Approval Menggunakan Algoritma Klasifikasi," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, hlm. 1342–1347, 2025.

[15] D. Athallah, F. Fathoni, dan M. I. F. Rachmad, "Klasifikasi Risiko Strok Menggunakan Algoritma Random Forest dengan Teknik Knowledge Discovery in Database," Indonesian Journal Computer Science, vol. 4, no. 1, hlm. 38-44, 2025.

[16] R. N. Mauliza dan Y. R. Sipayung, "Penerapan Text Mining Dalam Menganalisis Pendapat Masyarakat Terhadap Pemilu 2024 Pada Media Sosial X Menggunakan Metode Naive Bayes," *Technomedia Journal*, vol. 9, no. 1, hlm. 1–16, 2024.

[17] C. N. Oktariana dan N. R. Oktadini, "Analisis Sentimen Ulasan Pengguna Aplikasi Tokopedia Menggunakan Algoritma Random Forest," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, 2025.

[18] H. Hajaroh, T. Suprapti, dan R. Narasati, "Implementasi Algoritma Naive Bayes Untuk Analisis Sentimen Ulasan Produk Makanan Dan Minuman Di Tokopedia," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, hlm. 111–118, 2024.

[19] M. Idris, A. Rifai, dan K. D. Tania, "Sentiment Analysis of Tokopedia App Reviews using Machine Learning and Word Embeddings," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 9, no. 1, hlm. 210–219, 2025.