

ANALISIS USER PERSONA BERDASARKAN ULASAN PENGGUNA TOKOPEDIA MENGGUNAKAN METODE SVM BERBASIS WEB

Novemtri Sinulingga^{1*}, Asrar Aspia Manurung²

^{1,2} Universitas Muhammadiyah Sumatera Utara

*Penulis Korespondensi: sinulinggakrn@gmail.com

Abstract. *This research aims to build a web-based user persona analysis system that integrates sentiment classification using the Support Vector Machine (SVM) method with user review data of the Tokopedia application from the Google Play Store. A total of 5,000 reviews were collected through scraping techniques using the google-play-scraper library, then went through a cleaning process to become 4,781 valid reviews. The preprocessing stage includes cleaning, case folding, slang word normalization, tokenization, stopword removal, and stemming using the Sastrawi library, before being represented in numerical vector form using TF-IDF which produces 3,739 active features. Three-class sentiment classification (Positive, Neutral, Negative) was carried out using SVM with RBF kernel and class_weight='balanced'. Test results on 957 test data showed an accuracy of 81.09%, precision of 65.53%, recall of 60.18%, and F1-score of 61.35%, with an average K-Fold Cross Validation (K=5) of 81.46% ± 0.18%. The sentiment classification results are then processed using a keyword-scoring algorithm to form four user persona segments, namely Loyal Buyer (35.7%), Price Hunter (3.0%), Casual Shopper (4.7%), and Frustrated User (56.6%). All analysis results are visualized through a web-based dashboard built using Flask, MySQL, and HTML/CSS/JavaScript. This research proves that the integration of SVM and keyword-scoring algorithms is effective in producing meaningful user persona profiles and can be used as a basis for business strategy decision making and e-commerce platform product development.*

Keywords: *User Persona, Sentiment Analysis, Support Vector Machine, TF-IDF, Tokopedia*

Abstrak. Penelitian ini bertujuan membangun sistem analisis user persona berbasis web yang mengintegrasikan klasifikasi sentimen menggunakan metode Support Vector Machine (SVM) dengan data ulasan pengguna aplikasi Tokopedia dari Google Play Store. Sebanyak 5.000 ulasan dikumpulkan melalui teknik scraping menggunakan pustaka google-play-scraper, kemudian melalui proses pembersihan menjadi 4.781 ulasan valid. Tahap preprocessing meliputi cleaning, case folding, normalisasi kata slang, tokenisasi, stopword removal, dan stemming menggunakan pustaka Sastrawi, sebelum direpresentasikan dalam bentuk vektor numerik menggunakan TF-IDF yang menghasilkan 3.739 fitur aktif. Klasifikasi sentimen tiga kelas (Positif, Netral, Negatif) dilakukan menggunakan SVM dengan kernel RBF dan class_weight='balanced'. Hasil pengujian terhadap 957 data uji menunjukkan akurasi sebesar 81,09%, presisi 65,53%, recall 60,18%, dan F1-score 61,35%, dengan rata-rata K-Fold Cross Validation (K=5) sebesar 81,46% ± 0,18%. Hasil klasifikasi sentimen selanjutnya diolah menggunakan algoritma keyword-scoring untuk membentuk empat segmen user persona, yaitu Loyal Buyer (35,7%), Price Hunter (3,0%), Casual Shopper (4,7%), dan Frustrated User (56,6%). Seluruh hasil analisis divisualisasikan melalui dashboard berbasis web yang dibangun menggunakan Flask, MySQL, dan HTML/CSS/JavaScript. Penelitian ini membuktikan bahwa integrasi SVM dan algoritma keyword-scoring efektif menghasilkan profil persona pengguna yang bermakna dan dapat dimanfaatkan sebagai dasar pengambilan keputusan strategi bisnis serta pengembangan produk platform e-commerce.

Kata kunci: *User Persona, Analisis Sentimen, Support Vector Machine, TF-IDF, Tokopedia*

1. LATAR BELAKANG

Perkembangan teknologi informasi telah mengubah secara signifikan cara masyarakat menjalankan aktivitas belanja mereka. Era digital yang ditandai dengan

meluasnya akses internet mendorong pertumbuhan pesat industri e-commerce di Indonesia. Berdasarkan data Asosiasi Penyelenggara Jasa Internet Indonesia (APJII, 2024), lebih dari 215 juta penduduk Indonesia kini telah terhubung ke internet, dengan mayoritas di antaranya mengandalkan platform belanja daring sebagai solusi utama untuk memenuhi kebutuhan sehari-hari. Tokopedia, yang didirikan pada 17 Agustus 2009 oleh William Tanuwijaya dan Leontinus Alpha Edison, telah menjadi salah satu platform e-commerce terkemuka di Indonesia. Dengan berhasil meraih status unicorn, Tokopedia memfasilitasi jutaan transaksi setiap harinya melalui beragam fitur andal, seperti TopAds, Tokopedia Play, Flash Sale, serta layanan logistik terpadu yang bermitra dengan penyedia pengiriman seperti JNE dan SiCepat.

Peningkatan volume aktivitas transaksi ini telah menghasilkan akumulasi data ulasan pengguna dalam jumlah yang sangat besar. Diproyeksikan bahwa pada tahun 2025, Tokopedia akan menerima sekitar 50 juta ulasan pengguna setiap tahunnya, yang sebagian besar berasal dari platform seperti Google Play Store (Demetria & Wedhasmara, 2025). Ulasan-ulasan tersebut merepresentasikan pengalaman nyata jutaan konsumen dan mengandung potensi informasi yang sangat penting untuk mendukung pengembangan strategi bisnis platform. Meskipun demikian, besarnya volume data ulasan yang tersedia ternyata belum dimanfaatkan secara optimal. Riset menunjukkan bahwa ulasan pengguna memberikan pengaruh yang sangat besar pada ekosistem bisnis digital. Ulasan dari pengguna menyumbang sekitar 33,9% hingga 58,2% dalam memengaruhi proses pengambilan keputusan pembelian konsumen. Ulasan positif memiliki potensi untuk meningkatkan tingkat konversi penjualan hingga tiga kali lipat, sebaliknya apabila ulasan negatif mencapai lebih dari 25%, hal ini dapat memicu penurunan penjualan yang signifikan (Pratama & Wibowo, 2025).

Banyak penelitian telah dilakukan untuk menganalisis ulasan pengguna Tokopedia, tetapi sebagian besar masih fokus pada analisis sentimen saja tanpa menghasilkan profil user persona yang dapat langsung dimanfaatkan. Salsabila et al. (2024) melakukan analisis sentimen terhadap ulasan Tokopedia di Google Play Store dengan menerapkan algoritma Naïve Bayes yang dibandingkan dengan Support Vector Machine (SVM) dan Logistic Regression, dan diperoleh kesimpulan bahwa SVM menunjukkan performa terbaik. Setiawan dan Lumban (2024) melakukan perbandingan kinerja algoritma SVM

dan Naïve Bayes dalam analisis data ulasan e-commerce, serta mengonfirmasi superioritas SVM dibandingkan Naïve Bayes. Dalam studi lebih lanjut, Fiddin et al. (2025) memanfaatkan SVM yang dikombinasikan dengan representasi TF-IDF untuk analisis sentimen ulasan produk sayur di platform Tokopedia, menghasilkan tingkat akurasi sebesar 98%. Meskipun kedua penelitian ini berhasil mengembangkan model klasifikasi sentimen dengan performa tinggi, keduanya menghadapi keterbatasan mendasar yang serupa, yaitu model yang dihasilkan hanya mampu mengkategorikan ulasan ke dalam label sentimen positif atau negatif tanpa memberikan wawasan yang lebih strategis mengenai identitas pengguna maupun kebutuhan spesifik mereka.

User persona, yang didefinisikan sebagai representasi semi-fiktif dari segmen pengguna nyata yang dibangun berdasarkan data riset perilaku dan preferensi (Amplitude, 2023), merupakan pendekatan yang jauh lebih komprehensif dan dapat ditindaklanjuti secara langsung oleh tim pengembang maupun pemasaran. Menurut Interaction Design Foundation (2025), user persona adalah representasi yang menggabungkan tujuan, pola perilaku, dan interaksi pengguna ke dalam sebuah profil terstruktur yang berfungsi sebagai acuan untuk merancang solusi yang lebih efektif. Salah satu celah penelitian yang berhasil diidentifikasi adalah belum adanya studi komprehensif yang secara integratif menganalisis user persona berdasarkan ulasan pengguna di platform Tokopedia dengan menerapkan metode machine learning, serta memvisualisasikan hasil tersebut melalui sistem berbasis web.

Penelitian ini hadir untuk menjawab celah tersebut dengan membangun sebuah sistem yang secara otomatis menganalisis user persona berdasarkan ulasan pengguna Tokopedia menggunakan metode Support Vector Machine (SVM) yang diintegrasikan dengan pipeline Natural Language Processing (NLP). NLP berperan sebagai penghubung antara bahasa manusia dengan sistem pemrosesan digital, sehingga teknologi kecerdasan buatan dapat menangkap makna secara kontekstual, bukan sekadar mendeteksi kata kunci tertentu (Jurafsky & Martin, 2021). Metodologi yang diusulkan terdiri dari empat tahapan utama, yaitu preprocessing teks, representasi fitur menggunakan TF-IDF, klasifikasi menggunakan SVM, serta penyajian hasil dalam bentuk dashboard interaktif berbasis web. Penelitian ini menjadi semakin mendesak mengingat tingkat persaingan di industri e-commerce Indonesia yang kian meningkat. Tanpa pemahaman mendalam yang

didukung oleh data terkait karakteristik pengguna, platform e-commerce berpotensi membuat kesalahan dalam merancang fitur maupun strategi pemasaran yang kurang efektif. Berdasarkan rumusan masalah tersebut, penelitian ini bertujuan untuk: (1) mengimplementasikan preprocessing dan representasi data ulasan Tokopedia sebagai masukan klasifikasi SVM; (2) mengevaluasi performa SVM dalam klasifikasi sentimen tiga kelas berdasarkan akurasi, presisi, recall, dan F1-score; serta (3) menganalisis pola user persona dari hasil klasifikasi sentimen dan membangun dashboard berbasis web untuk visualisasi profil persona, tren sentimen, dan pain point pengguna.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis komputasi dengan memanfaatkan metode text mining dan teknik machine learning untuk mengklasifikasikan sentimen ulasan pengguna Tokopedia sekaligus membentuk profil user persona.

1. Sumber dan Jenis Data

Data yang digunakan merupakan data primer berupa ulasan pengguna aplikasi Tokopedia (com.tokopedia.tkpdp) yang tersedia secara publik di Google Play Store. Pengumpulan data dilakukan dengan teknik scraping menggunakan pustaka google-play-scraper berbasis Python melalui Google Colaboratory, dengan parameter lang='id', country='id', dan Sort=NEWEST untuk memperoleh ulasan terbaru berbahasa Indonesia. Proses scraping berhasil mengumpulkan 5.000 ulasan dengan periode Maret hingga Juni 2026, mencakup delapan atribut yaitu nomor urut, id ulasan, nama pengguna, teks ulasan, rating, tanggal ulasan, jumlah likes, dan versi aplikasi. Data disimpan dalam format Microsoft Excel (.xlsx), kemudian melalui tahap pembersihan (penghapusan data duplikat, baris kosong, dan rating tidak valid) sehingga menjadi 4.781 ulasan valid yang digunakan dalam proses analisis.

2. Variabel Penelitian

Variabel utama yang digunakan dalam proses klasifikasi terdiri dari:

- a. Teks ulasan (teks_ulasan), sebagai fitur utama yang direpresentasikan menggunakan TF-IDF

- b. Rating bintang (rating) skala 1–5, sebagai dasar penentuan label sentimen (Positif, Netral, Negatif)

Selain itu, terdapat variabel pendukung berupa skor kata kunci (skor harga, skor positif, skor layanan) yang tidak digunakan dalam proses klasifikasi SVM, namun dimanfaatkan sebagai dasar algoritma keyword-scoring dalam pembentukan user persona. Skor harga dihitung dari kemunculan kata seperti 'murah', 'promo', 'diskon', 'voucher', dan 'cashback'; skor positif dari kata seperti 'bagus', 'mudah', 'aman', 'puas', dan 'cepat'; sedangkan skor layanan dari kata seperti 'kirim', 'kurir', 'cs', dan 'refund'.

3. Tahapan Pengolahan Data

a. Text Preprocessing

Data ulasan mentah melewati enam tahap preprocessing yang dijalankan secara berurutan: (1) cleaning, yaitu penghapusan URL dan karakter non-alfabet menggunakan regular expression; (2) case folding, yaitu mengubah seluruh huruf menjadi huruf kecil untuk menjaga konsistensi; (3) normalisasi kata slang menggunakan kamus kustom bahasa Indonesia, misalnya kata 'gk' menjadi 'tidak', 'bgt' menjadi 'sangat', dan 'cs' menjadi 'customer service'; (4) tokenisasi menggunakan `word_tokenize()` dari pustaka NLTK; (5) stopwords removal menggunakan gabungan daftar stopwords NLTK bahasa Indonesia dan daftar tambahan kustom seperti 'tokopedia', 'toped', dan 'yg'; serta (6) stemming menggunakan pustaka Sastrawi untuk mengubah kata berimbuhan ke bentuk dasarnya.

b. Pelabelan Sentimen

Pelabelan sentimen dilakukan dengan pendekatan rule-based berdasarkan nilai rating bintang, di mana rating 4–5 bintang diklasifikasikan sebagai sentimen Positif, rating 3 bintang sebagai Netral, dan rating 1–2 bintang sebagai Negatif. Dari 4.781 data yang diberi label, diperoleh distribusi 1.962 ulasan Positif (41,0%), 283 ulasan Netral (5,9%), dan 2.536 ulasan Negatif (53,0%).

c. Representasi Fitur TF-IDF

Teks yang telah diproses direpresentasikan dalam bentuk vektor numerik menggunakan `TfidfVectorizer` dari `scikit-learn` dengan konfigurasi

max_features=5.000, ngram_range=(1,2) untuk mempertimbangkan unigram dan bigram, min_df adaptif, max_df=0,95, dan sublinear_tf=True untuk transformasi logaritmik pada nilai TF, sehingga menghasilkan 3.739 fitur aktif.

d. Proses Klasifikasi SVM

Data dibagi menjadi 80% data latih (3.824 data) dan 20% data uji (957 data) menggunakan stratified train_test_split dengan random_state=42 agar hasil dapat direproduksi. Model klasifikasi dibangun menggunakan SVC dengan kernel Radial Basis Function (RBF), parameter C=1,0, gamma='scale' (dihitung otomatis sebagai 1 dibagi jumlah fitur dikali varians data), class_weight='balanced' untuk mengatasi ketidakseimbangan distribusi kelas, dan probability=True untuk mengaktifkan estimasi probabilitas melalui Platt scaling. Strategi multi-kelas yang digunakan adalah One-vs-One, di mana setiap pasangan kelas dibandingkan secara biner dan keputusan akhir ditentukan melalui voting.

e. Evaluasi Model

Evaluasi dilakukan menggunakan confusion matrix serta empat metrik utama, yaitu akurasi (accuracy_score), presisi (precision_score, average='macro'), recall (recall_score, average='macro'), dan F1-score (f1_score, average='macro'), dengan parameter zero_division=0 untuk menangani kelas tanpa prediksi. Validasi tambahan dilakukan melalui Stratified K-Fold Cross Validation dengan K=5 untuk menilai stabilitas dan konsistensi model terhadap pembagian data yang berbeda-beda, guna memastikan model tidak mengalami overfitting terhadap data latih.

4. Pembentukan User Persona

Segmen user persona dibentuk melalui algoritma keyword-scoring yang menghitung tiga jenis skor kata kunci pada setiap ulasan, yaitu skor harga, skor positif, dan skor layanan. Penentuan segmen dilakukan berdasarkan kombinasi label sentimen hasil SVM dengan dominasi skor kata kunci sebagai berikut: (1) Loyal Buyer, yaitu ulasan bersentimen Positif yang tidak didominasi skor harga; (2) Price Hunter, yaitu ulasan bersentimen Positif dengan skor harga mendominasi skor positif dan skor layanan, atau ulasan bersentimen Netral dengan skor harga lebih dari nol; (3) Casual Shopper, yaitu ulasan bersentimen Netral dengan skor harga sama dengan nol; dan (4)

Frustrated User, yaitu seluruh ulasan bersentimen Negatif tanpa memandang topik yang diulas.

5. Pengembangan Sistem Berbasis Web

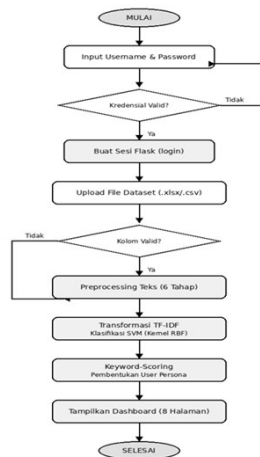
Seluruh pipeline analisis mulai dari preprocessing, klasifikasi SVM, hingga pembentukan user persona diintegrasikan ke dalam sistem berbasis web menggunakan Flask sebagai backend Python dan HTML/CSS/JavaScript sebagai frontend, dengan tampilan dashboard interaktif yang terdiri dari delapan halaman: Dashboard Utama, Klasifikasi SVM, Persona Pengguna, Analisis Sentimen, Dataset Ulasan, Word Cloud, Upload Dataset, dan Prediksi Teks Baru. Sistem juga dilengkapi dengan autentikasi pengguna berbasis MySQL dan manajemen sesi persisten menggunakan format Parquet untuk dataset, Pickle untuk model SVM dan TF-IDF vectorizer, serta JSON untuk metadata evaluasi.

6. Pengujian dan Evaluasi Sistem

Pengujian sistem dilakukan pada dua aspek utama, yaitu pengujian kinerja model klasifikasi SVM menggunakan empat metrik evaluasi dan K-Fold Cross Validation, serta pengujian fungsionalitas sistem menggunakan metode blackbox testing terhadap seluruh use case yang telah dirancang, mencakup skenario normal (happy path) maupun skenario kesalahan (error path).

Alur Kerja Sistem

Sistem analisis user persona berbasis SVM dirancang sebagai sistem pendukung analisis yang membantu peneliti maupun pelaku bisnis dalam memahami karakteristik pengguna Tokopedia berdasarkan ulasan yang mereka tulis. Sistem menerima data ulasan hasil scraping sebagai input utama, kemudian diolah melalui pipeline preprocessing, klasifikasi SVM, dan algoritma keyword-scoring untuk menghasilkan output berupa label sentimen dan profil user persona yang divisualisasikan melalui dashboard interaktif berbasis web.



Gambar 1. Flowchart Alur Sistem

3. HASIL DAN PEMBAHASAN

1. Hasil Pengumpulan dan Preprocessing Data

Proses scraping menghasilkan 5.000 ulasan mentah dengan distribusi rating sebagai berikut: bintang 1 sebanyak 2.275 ulasan (45,5%), bintang 2 sebanyak 280 ulasan (5,6%), bintang 3 sebanyak 288 ulasan (5,8%), bintang 4 sebanyak 243 ulasan (4,9%), dan bintang 5 sebanyak 1.914 ulasan (38,3%). Setelah melalui tahap pembersihan, diperoleh 4.781 ulasan valid. Sebagai ilustrasi, teks ulasan asli seperti 'ok mantap aplikasi ini membantu bangtt' setelah melalui enam tahap preprocessing menjadi 'oke mantap aplikasi bantu bangtt', sedangkan ulasan 'voucher gak bisa dipakai, gratis ongkir itu mitos, toko fiktif, harga ga ngotak' menjadi 'voucher pakai gratis ongkir mitos toko fiktif harga ngotak'. Proses ini secara efektif menghilangkan noise seperti tanda baca berlebihan dan kata tidak baku, sekaligus mempertahankan makna inti dari ulasan.

2. Hasil Proses Klasifikasi SVM

Proses klasifikasi dilakukan menggunakan algoritma SVM dengan kernel RBF terhadap 4.781 data ulasan yang telah melalui tahap preprocessing dan representasi TF-IDF dengan 3.739 fitur aktif. Pengujian model terhadap 957 data uji menghasilkan performa sebagaimana dirangkum pada Tabel 1.

Metrik Evaluasi	Nilai	Keterangan
-----------------	-------	------------

Akurasi	81,09%	776/957 data uji benar
Presisi (Macro)	65,53%	Rata-rata per kelas
Recall (Macro)	60,18%	Rata-rata per kelas
F1-Score (Macro)	61,35%	Harmonic mean
K-Fold CV (K=5)	81,46% $\pm$ 0,18%	Model stabil, tidak overfitting

Tingkat akurasi sebesar 81,09% menunjukkan bahwa dari 957 data uji, sebanyak 776 ulasan berhasil diklasifikasikan dengan tepat. Rincian hasil klasifikasi pada tingkat kelas ditunjukkan melalui confusion matrix pada Tabel 2.

	Prediksi Positif	Prediksi Netral	Prediksi Negatif
Aktual Positif	290	10	93
Aktual Netral	7	7	43
Aktual Negatif	19	9	479

Berdasarkan Tabel 2, model SVM mampu mengklasifikasikan kelas Positif dengan benar pada 290 dari 393 data aktual (73,8%), kelas Netral pada 7 dari 57 data aktual (12,3%), dan kelas Negatif pada 479 dari 507 data aktual (94,5%). Nilai presisi dan recall yang lebih rendah dibandingkan akurasi disebabkan oleh ketimpangan distribusi kelas, terutama pada kelas Netral yang hanya berjumlah 283 data (5,9%) dari keseluruhan dataset. Meskipun parameter `class_weight='balanced'` telah diterapkan untuk menyeimbangkan bobot pelatihan, keterbatasan jumlah sampel pada kelas Netral tetap menjadi hambatan utama bagi model dalam mengenali pola yang membedakannya dari kelas lain.

Hasil validasi Stratified K-Fold Cross Validation dengan K=5 menghasilkan skor akurasi per fold sebesar 81,44%, 81,70%, 81,44%, 81,57%, dan 81,15%, dengan rata-rata

81,46% dan standar deviasi hanya $\pm 0,18\%$. Perbedaan antara fold tertinggi dan terendah hanya sebesar 0,55%, yang mengindikasikan bahwa model memiliki stabilitas tinggi, tidak bergantung pada pembagian data tertentu, dan tidak mengalami overfitting terhadap data latih.

3. Karakteristik / Distribusi User Persona

Hasil pembentukan user persona terhadap 4.781 ulasan menghasilkan empat segmen sebagaimana disajikan pada Tabel 3.

Segmen Persona	Jumlah Ulasan	Persentase	Dasar Sentimen
Loyal Buyer	1.709	35,7%	Positif
Frustrated User	2.705	56,6%	Negatif
Casual Shopper	223	4,7%	Netral
Price Hunter	144	3,0%	Positif/Netral

Loyal Buyer merupakan kelompok pengguna dengan sentimen positif yang tidak didominasi kata kunci harga, mencerminkan basis pengguna setia yang puas terhadap layanan Tokopedia secara keseluruhan tanpa terlalu mempertimbangkan faktor harga. Frustrated User mencakup seluruh ulasan bersentimen negatif tanpa memandang topik yang diulas, mencerminkan pengguna dengan pengalaman buruk terhadap aplikasi. Casual Shopper merupakan pengguna dengan sentimen netral tanpa preferensi harga tertentu, cenderung menggunakan Tokopedia secara tidak rutin. Sementara itu, Price Hunter merupakan pengguna yang secara eksplisit berorientasi pada harga, diskon, dan promosi dalam ulasannya.

4. Perbandingan dengan Penelitian Terkait

Dibandingkan dengan penelitian terkait yang menggunakan dataset serupa, akurasi 81,09% yang diperoleh berada sedikit di bawah hasil Fiddin et al. (2025) yang mencapai 98% pada dataset satu kategori produk dan Oktariana dan Oktadini (2025) yang mencapai 94,5% menggunakan algoritma Random Forest. Perbedaan ini wajar mengingat penelitian ini menggunakan klasifikasi tiga kelas (Positif, Netral, Negatif) yang secara

inheren lebih sulit dibandingkan klasifikasi biner, serta menggunakan dataset multi-kategori yang lebih beragam dengan 4.781 ulasan yang mencakup berbagai topik keluhan pengguna Tokopedia, bukan hanya satu kategori produk tertentu.

5. Implementasi Dashboard dan Pengujian Sistem

Seluruh hasil analisis divisualisasikan melalui dashboard berbasis web yang mengintegrasikan fitur upload dataset, prediksi teks baru, word cloud per kelas sentimen, filter dan eksplorasi dataset, serta ekspor hasil klasifikasi dalam format CSV. Pengujian fungsionalitas sistem menggunakan metode blackbox testing terhadap 18 skenario dari 13 use case, mencakup skenario login dengan kredensial valid maupun tidak valid, upload dataset dengan berbagai format, prediksi teks dengan input valid maupun kosong, filter dataset, hingga ekspor dan reset sesi, menunjukkan bahwa seluruh fitur inti sistem berfungsi sesuai rancangan baik pada skenario normal maupun skenario kesalahan.

6. Pembahasan

Dominasi segmen Frustrated User sebesar 56,6% mengindikasikan bahwa mayoritas pengguna yang menulis ulasan pada periode pengumpulan data adalah pengguna dengan pengalaman negatif, konsisten dengan fenomena negativity bias pada ulasan daring, di mana pengguna yang tidak puas cenderung lebih termotivasi untuk menulis ulasan dibandingkan pengguna yang puas (Pratama & Wibowo, 2025). Segmen Loyal Buyer menempati posisi kedua dengan 35,7%, mencerminkan basis pengguna setia yang dianggap sebagai salah satu aset paling berharga karena menunjukkan tingkat loyalitas organik yang tinggi terhadap platform. Sementara itu, segmen Price Hunter (3,0%) dan Casual Shopper (4,7%) memiliki porsi yang relatif kecil, namun tetap memberikan wawasan berharga terkait motivasi spesifik pengguna dalam memanfaatkan platform Tokopedia.

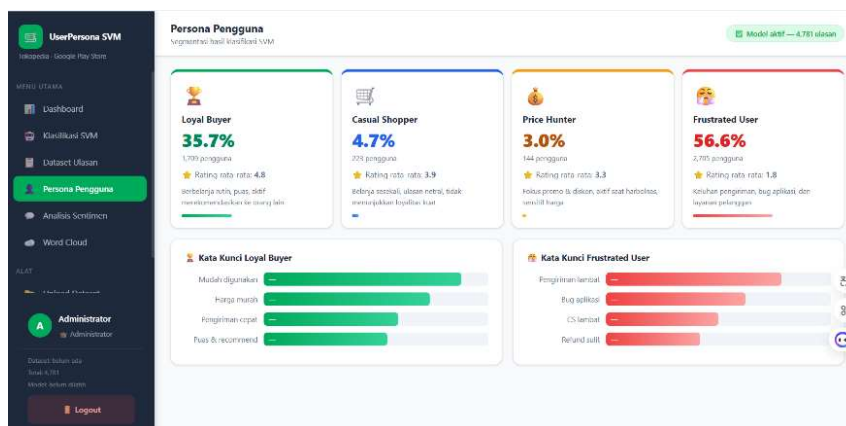
Temuan ini menunjukkan bahwa faktor pengalaman penggunaan dan kepercayaan memainkan peran penting dalam menentukan sentimen dan intensitas penggunaan platform. Oleh karena itu, penyedia layanan e-commerce perlu memperhatikan aspek kemudahan, kecepatan pengiriman, dan keamanan transaksi untuk meningkatkan kepuasan pengguna, khususnya dalam menekan proporsi segmen Frustrated User. Secara keseluruhan, integrasi SVM dengan algoritma keyword-scoring mampu mengidentifikasi

profil pengguna secara efektif dan dapat dimanfaatkan sebagai dasar pengambilan keputusan, baik dalam pengembangan fitur produk maupun perumusan strategi pemasaran pada sektor e-commerce di Indonesia.



Gambar 2. Tampilan Halaman Dashboard Utama

Gambar 2 menampilkan ringkasan hasil analisis secara menyeluruh, mencakup total ulasan, akurasi model, distribusi sentimen, dan distribusi user persona secara real-time. Tampilan ini menunjukkan bahwa sistem mampu mengintegrasikan seluruh pipeline analisis ke dalam satu antarmuka yang informatif.



Gambar 3. Tampilan Halaman Persona Pengguna

Gambar 3 menunjukkan hasil pengelompokan user persona yang ditampilkan sistem, meliputi jumlah ulasan, persentase, karakteristik, serta kata kunci dominan pada masing-masing segmen. Tampilan ini mempermudah pemangku kepentingan bisnis dalam memahami profil pengguna tanpa perlu menganalisis data mentah secara manual.

4. KESIMPULAN DAN SARAN

Kesimpulan :

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, dapat disimpulkan beberapa hal sebagai berikut. Pertama, proses preprocessing dan representasi data pada ulasan pengguna aplikasi Tokopedia telah berhasil dilakukan melalui pipeline enam tahap, yaitu cleaning, case folding, normalisasi kata slang, tokenisasi, stopwords removal, dan stemming. Data sebanyak 4.781 ulasan hasil scraping Google Play Store direpresentasikan menggunakan TF-IDF Vectorizer dengan konfigurasi `max_features=5.000` dan `ngram_range=(1,2)`, menghasilkan 3.739 fitur aktif yang siap digunakan sebagai input klasifikasi.

Kedua, model Support Vector Machine dengan kernel RBF terbukti efektif dalam mengklasifikasikan sentimen ulasan Tokopedia ke dalam tiga kategori: Positif, Netral, dan Negatif. Berdasarkan evaluasi terhadap 957 data uji, model ini mencapai akurasi 81,09%, dengan presisi sebesar 65,53%, recall 60,18%, dan F1-score 61,35%. Validasi K-Fold Cross Validation (K=5) menghasilkan rata-rata akurasi sebesar $81,46\% \pm 0,18\%$, menunjukkan bahwa model tidak mengalami overfitting dan memiliki kemampuan generalisasi yang baik, meskipun performa presisi dan recall masih terkendala oleh ketidakseimbangan jumlah data pada kelas Netral.

Ketiga, pembentukan user persona berhasil dilakukan melalui algoritma keyword-scoring yang mengintegrasikan hasil klasifikasi SVM dengan skor dari tiga kelompok kata kunci. Sistem ini menghasilkan empat segmen pengguna, yaitu Loyal Buyer dengan 1.709 ulasan (35,7%), Price Hunter dengan 144 ulasan (3,0%), Casual Shopper dengan 223 ulasan (4,7%), dan Frustrated User dengan 2.705 ulasan (56,6%). Tingginya proporsi segmen Frustrated User mencerminkan besarnya tingkat ketidakpuasan dalam dataset yang digunakan pada periode penelitian.

Keempat, seluruh hasil analisis berhasil divisualisasikan melalui sistem antarmuka berbasis web yang dikembangkan menggunakan Flask, MySQL, serta kerangka HTML/CSS/JavaScript, mencakup delapan halaman dashboard interaktif. Pengujian blackbox terhadap 18 skenario dari 13 use case menunjukkan bahwa seluruh fitur inti sistem berjalan sesuai dengan yang diharapkan. Dengan demikian, penelitian ini

membuktikan bahwa integrasi SVM dengan algoritma keyword-scoring efektif menghasilkan profil pengguna yang bermakna dan dapat dimanfaatkan sebagai dasar pengambilan keputusan strategi bisnis serta pengembangan produk pada platform e-commerce.

Saran :

1. Pelabelan sentimen dengan pendekatan rule-based yang didasarkan pada rating bintang sering kali tidak sepenuhnya merepresentasikan sentimen sebenarnya dari teks ulasan. Penelitian selanjutnya dapat mempertimbangkan penerapan pelabelan manual oleh ahli (human annotation) atau metode semi-supervised learning, khususnya untuk meningkatkan kualitas pelabelan pada kelas Netral.
2. Kelas Netral hanya terdiri dari 283 ulasan (5,9%), yang mengakibatkan penurunan nilai presisi dan recall. Penelitian mendatang disarankan untuk mengeksplorasi berbagai teknik augmentasi data guna mengatasi ketidakseimbangan kelas yang signifikan tersebut.
3. Perbandingan performa SVM dengan metode lain seperti Naive Bayes, Random Forest, atau model berbasis transformer seperti IndoBERT dapat dilakukan untuk menemukan model yang lebih optimal dalam tugas klasifikasi teks ulasan berbahasa Indonesia informal.
4. Sistem yang ada saat ini masih mengandalkan model SVM statis yang tidak diperbarui secara otomatis. Pengembangan di masa mendatang dapat mencakup penambahan fitur penjadwalan pelatihan ulang secara berkala, sehingga model dapat terus menyesuaikan diri dengan pola sentimen pengguna yang terbaru.
5. Sistem dapat ditingkatkan dengan mengintegrasikan fitur analisis berbasis aspek (aspect-based sentiment analysis) yang mampu mendeteksi sentimen terhadap aspek tertentu, seperti harga, pengiriman, dan layanan pelanggan secara lebih terperinci, sehingga menghasilkan profil persona pengguna yang lebih mendalam

DAFTAR REFERENSI

- Amplitude. (2023). User Persona: Definisi dan Penerapannya dalam Riset Pengguna.
- APJII. (2024). Survei Penetrasi dan Perilaku Pengguna Internet Indonesia 2024. Asosiasi Penyelenggara Jasa Internet Indonesia.

- Demetria, R., & Wedhasmara, A. (2025). Proyeksi volume ulasan pengguna platform e-commerce di Indonesia.
- Fiddin, F., et al. (2025). Analisis Sentimen Ulasan Produk Sayur di Tokopedia Menggunakan Model Support Vector Machine dengan Representasi TF-IDF. *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)*, 24(2), 162–172.
- Hendrawan, I. R. (2023). *Natural Language Processing*. Yogyakarta: CV Andi Offset.
- Huda, D. N. (2023). Analisis Sentimen Layanan Jasa Pengiriman pada Ulasan Play Store: Systematic Literature Review. *Jurnal Informatika dan Teknologi Komputer*, 4(2), 87–98.
- Interaction Design Foundation. (2025). What are Personas?
- Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.).
- Lestari, T. P. (2022). Analisis Text Mining pada Sosial Media Twitter Menggunakan Metode Support Vector Machine (SVM) dan Social Network Analysis (SNA). *Jurnal Informatika Ekonomi Bisnis*, 4(3), 65–71.
- Milal, I. S. (2023). Klasifikasi Teks Review pada E-commerce Tokopedia Menggunakan Algoritma SVM. *Jurnal Ilmiah Nasional Riset Aplikasi dan Teknik Informatika*, 5(1), 34–45.
- Munthe, S. R. (2025). Pemodelan Sentimen dalam Ulasan Produk dan Penerapannya pada E-commerce.
- Nielsen Norman Group. (2025). *Personas: Study Guide*.
- Oktariana, C. N., & Oktadini, N. (2025). Analisis Sentimen Ulasan Pengguna Aplikasi Tokopedia Menggunakan Algoritma Random Forest. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(6), 1224–1231.
- Pratama, D., & Wibowo, A. (2025). Pengaruh Ulasan Pengguna terhadap Keputusan Pembelian pada Platform E-commerce.
- Salsabila, N. A., et al. (2024). Analisis Sentimen pada Ulasan Aplikasi Tokopedia Menggunakan Klasifikasi Naïve Bayes. *PRISMA, Prosiding Seminar Nasional Matematika*, 7, 44–51.
- Setiawan, T., et al. (2024). Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Komentar TikTok pada Produk Skincare. *AICOMS: Applied Information Technology and Computer Science*, 3(2), 28–32.
- Vapnik, V., et al. (1992). A Training Algorithm for Optimal Margin Classifiers.
- Wardana, A. K. (2025). Analisis Sentimen Ulasan Produk E-Commerce Menggunakan Naive Bayes dan TF-IDF. *Jurnal Ilmiah Multidisiplin Ilmu*, 1(1), 63–77.
- Wijaya, A. C. (2024). Perancangan Sistem Dashboard Penjualan Berbasis Web untuk Toko Online Caro Cara. *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, 13(3), 2114–2125.
- Yaniasih, & Budi, I. (2025). *Analisis Sentimen Berdasarkan Data Media Sosial*. Depok: UI Publishing.