ANTHOR
Education and Learning Journal
https://anthor.org/index.php/anthor
Email. anthor.Jurnal@gmail.com
E-ISSN: 2963 –198X P-ISSN: 2963–2498 | Halaman: 1690 - 1700
Volume 5 Nomor 5 Tahun 2026

Analisis Sentimen Ulasan Pengguna terhadap Produk Skincare Sephora Menggunakan Algoritma Logistic Regression

Informasi Naskah

Ulfah Ulfiah

Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer, Indonesia

ulfiahulfah67@gmail.com

Abstrak

Banyaknya ulasan produk perawatan kulit di platform digital menimbulkan tantangan dalam mengolah informasi opini konsumen secara manual dan efisien. Penelitian ini bertujuan mengklasifikasikan sentimen ulasan pengguna terhadap produk skincare Sephora ke dalam tiga kelas, yaitu positif, netral, dan negatif. Metode yang digunakan adalah kuantitatif eksperimental dengan mengombinasikan ekstraksi fitur Term Frequency–Inverse Document Frequency (TF-IDF) dan algoritma Logistic Regression. Pengolahan data mencakup pembersihan teks, penanganan duplikasi, eliminasi konflik label, serta penyeimbangan bobot kelas (class_weight = 'balanced'). Dari total 245.092 data bersih yang dibagi dengan rasio 80:20, model dievaluasi pada 49.019 data uji. Hasil pengujian menunjukkan bahwa model mencapai akurasi sebesar 86,00%, macro-F1 sebesar 0,6778, dan weighted-F1 sebesar 0,8752. Berdasarkan evaluasi per kelas, sentimen positif memperoleh F1-score tertinggi sebesar 0,938, sentimen negatif sebesar 0,687, sedangkan sentimen netral sebesar 0,409. Kesimpulannya, kombinasi TF-IDF dan Logistic Regression efektif dalam memprediksi sentimen berbasis rating, terutama pada kelas positif, namun memiliki keterbatasan pada kelas netral pernyataan yang lebih hati-hati penyebab pasti rendahnya F1-score netral karena hasil eksperimen belum membuktikan hubungan sebab-akibat secara langsung.

Kata Kunci:
Analisis Sentimen, Logistic Regression, Skincare Sephora, TF-IDF, Ulasan Pengguna.

Abstract

The large number of skincare product reviews on digital platforms presents challenges in processing consumer opinions manually and efficiently. This study aims to classify user reviews of Sephora skincare products into three sentiment categories: positive, neutral, and negative. An experimental quantitative approach was employed by combining Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction with the Logistic Regression algorithm. Data processing involved text cleaning, duplicate handling, the removal of conflicting labels, and class weight balancing (`class_weight = 'balanced'`). Of the 245,092 cleaned reviews, 80% were allocated to the training set and 20% to the test set, resulting in 49,019 test samples. The evaluation results showed that the model achieved an accuracy of 86.00%, a macro-F1 score of 0.6778, and a weighted-F1 score of 0.8752. In the per-class evaluation, positive sentiment achieved the highest F1-score of 0.938, followed by negative sentiment at 0.687, while neutral sentiment achieved an F1-score of 0.409. In conclusion, the combination of TF-IDF and Logistic Regression demonstrated good performance in predicting rating-based sentiment, particularly for positive reviews, but showed limitations in classifying neutral reviews. Class imbalance and potential ambiguity in review language may be associated with the lower F1-score for neutral sentiment. However, the experimental results do not establish a direct causal relationship between these factors and the model's lower performance on the neutral class.

***Keyword**:*
User Reviews, Logistic Regression, Sephora Skincare, Sentiment Analysis, TF-IDF.

Information Production:
© 2026 *by the authors of this article. This work is licensed under an Attribution-ShareAlike* 4.0 *International* (CC BY-SA 4.0)

PENDAHULUAN

Perkembangan teknologi informasi telah mendorong pertumbuhan perdagangan digital yang memungkinkan konsumen memperoleh informasi produk, membandingkan alternatif, dan menyampaikan pengalaman penggunaan melalui ulasan daring. Dalam industri perawatan kulit skincare, ulasan konsumen menjadi salah satu sumber informasi yang dapat membantu calon pembeli memahami pengalaman pengguna lain terhadap suatu produk. Ulasan tersebut dapat mencakup berbagai aspek, seperti tekstur, aroma, kenyamanan penggunaan, kemasan, harga, dan persepsi terhadap manfaat produk. Selain berfungsi sebagai sarana berbagi pengalaman, ulasan konsumen juga menyediakan data yang dapat dimanfaatkan untuk memahami kecenderungan opini pengguna terhadap produk.

Meskipun demikian, banyaknya ulasan yang tersedia pada platform digital menimbulkan tantangan dalam pengolahan informasi. Ulasan ditulis menggunakan bahasa alami dengan variasi kosakata, gaya bahasa

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

informal, singkatan, dan ekspresi yang dapat mengandung penilaian berbeda. Sebagian ulasan menunjukkan kepuasan terhadap produk, sebagian lainnya menyampaikan ketidakpuasan, sedangkan ulasan tertentu memiliki kecenderungan netral. Keragaman tersebut membuat analisis manual menjadi kurang efisien ketika jumlah data besar. Selain itu, rating yang diberikan pengguna tidak selalu sepenuhnya mencerminkan sentimen yang terkandung dalam teks ulasan. Oleh karena itu, diperlukan pendekatan komputasional yang dapat mengolah ulasan secara sistematis dan mengevaluasi hasil klasifikasi secara terukur. Analisis sentimen merupakan salah satu pendekatan dalam pemrosesan bahasa alami Natural Language Processing atau NLP yang digunakan untuk mengidentifikasi dan mengelompokkan opini berdasarkan kecenderungan sentimennya. (Tan et al., 2023) membahas berbagai pendekatan, dataset, dan arah pengembangan penelitian analisis sentimen, sedangkan (Ligthart et al., 2021) mengkaji perkembangan penelitian sekunder di bidang tersebut serta berbagai tantangan metodologisnya. Penerapan analisis sentimen pada ulasan produk memungkinkan informasi berbentuk teks dirangkum ke dalam kategori yang lebih terstruktur sehingga kecenderungan opini konsumen dapat dianalisis secara lebih sistematis.

Salah satu algoritma machine learning yang dapat digunakan untuk klasifikasi teks adalah Logistic Regression. Algoritma ini memperkirakan probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan fitur yang digunakan. Dalam klasifikasi teks, Logistic Regression dapat dikombinasikan dengan Term Frequency–Inverse Document Frequency (TF-IDF), yaitu metode representasi teks yang memberikan bobot pada istilah berdasarkan frekuensi kemunculannya dalam dokumen dan tingkat penyebarannya di dalam korpus. Representasi TF-IDF memungkinkan teks diubah menjadi fitur numerik yang dapat digunakan oleh algoritma klasifikasi untuk mempelajari pola hubungan antara istilah dan label sentimen.

Penelitian (Fadly et al., 2022) mengkaji analisis sentimen pada 10.000 ulasan produk natural skincare dari Sephora.com dengan membandingkan beberapa pendekatan, termasuk Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Tree, dan Deep Learning. Penelitian tersebut memberikan konteks empiris mengenai penerapan klasifikasi sentimen pada ulasan produk perawatan kulit. Sementara itu, (Dikusuma & Widjaja, 2025) membandingkan Naïve Bayes dan Logistic Regression pada ulasan produk skincare di Tokopedia menggunakan TF-IDF. Kedua penelitian tersebut relevan sebagai landasan pembanding, tetapi hasilnya tidak dapat dibandingkan secara langsung tanpa mempertimbangkan perbedaan sumber data, skema pelabelan, distribusi kelas, dan rancangan eksperimen.

Salah satu tantangan penting dalam klasifikasi sentimen adalah ketidakseimbangan distribusi kelas. Ketika jumlah ulasan positif jauh lebih besar daripada jumlah ulasan netral dan negatif, nilai akurasi yang tinggi belum tentu menunjukkan bahwa model mampu mengenali setiap kelas secara seimbang. Pada dataset yang digunakan dalam penelitian ini, distribusi sebelum audit duplikasi dan penanganan konflik label terdiri atas 236.191 ulasan positif (82,85%), 21.148 ulasan netral (7,42%), dan 27.728 ulasan negatif (9,73%), dari total 285.067 ulasan setelah 345 data tanpa teks ulasan atau rating dikeluarkan. Distribusi tersebut menunjukkan dominasi kelas positif dan mengindikasikan perlunya evaluasi yang tidak hanya berfokus pada akurasi, tetapi juga memperhatikan precision, recall, F1-score setiap kelas, macro-F1, dan weighted-F1.

Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma Logistic Regression dengan representasi fitur TF-IDF untuk mengklasifikasikan ulasan pengguna terhadap produk skincare Sephora ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif. Data diproses melalui tahapan pembersihan teks, pemeriksaan kualitas data, penanganan duplikasi, dan penanganan konflik label. Label sentimen ditentukan berdasarkan rating pengguna, sehingga hasil klasifikasi perlu dipahami sebagai prediksi terhadap label berbasis rating, bukan sebagai hasil anotasi sentimen manual yang sepenuhnya didasarkan pada makna teks.

Fokus penelitian ini terletak pada penerapan dan evaluasi Logistic Regression untuk klasifikasi tiga kelas sentimen pada dataset ulasan Sephora dengan memperhatikan ketidakseimbangan kelas dan penanganan kualitas data. Kinerja model dievaluasi menggunakan akurasi, precision, recall, F1-score, macro-F1, dan weighted-F1. Penggunaan metrik tersebut bertujuan memberikan gambaran yang lebih menyeluruh mengenai kemampuan model dalam mengenali kelas mayoritas maupun kelas minoritas. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi empiris terhadap penerapan klasifikasi sentimen pada ulasan produk perawatan kulit serta menjadi referensi bagi penelitian selanjutnya yang mengembangkan metode klasifikasi teks dan evaluasi model pada data dengan distribusi kelas tidak seimbang.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental yang bertujuan untuk menerapkan dan mengevaluasi algoritma Logistic Regression dalam mengklasifikasikan sentimen ulasan pengguna terhadap produk skincare Sephora. Pendekatan kuantitatif dipilih karena proses evaluasi dilakukan berdasarkan hasil pengukuran kinerja model menggunakan metrik klasifikasi, yaitu accuracy, precision, recall,

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

F1-score, macro-F1, dan weighted-F1. Melalui pengukuran tersebut, kemampuan model dalam mengklasifikasikan ulasan ke dalam kategori sentimen positif, netral, dan negatif dapat dianalisis secara objektif.

Eksperimen dilakukan dengan mengubah data ulasan yang berbentuk teks tidak terstruktur menjadi representasi numerik menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Representasi numerik tersebut kemudian digunakan sebagai masukan bagi algoritma Logistic Regression untuk mempelajari hubungan antara karakteristik kata dalam ulasan dan label sentimen yang telah ditentukan berdasarkan rating pengguna.

Fokus penelitian ini bukan membandingkan beberapa algoritma klasifikasi, melainkan mengevaluasi kemampuan Logistic Regression ketika diterapkan pada dataset ulasan skincare dengan distribusi kelas yang tidak seimbang. Oleh karena itu, evaluasi tidak hanya berfokus pada tingkat akurasi keseluruhan, tetapi juga memperhatikan kemampuan model dalam mengenali masing-masing kelas sentimen, khususnya kelas netral yang memiliki jumlah data lebih sedikit dibandingkan kelas positif.

Secara keseluruhan, penelitian dilaksanakan melalui beberapa tahapan, meliputi pengumpulan dataset, pemeriksaan kualitas data, penentuan label sentimen, prapemrosesan teks, penghapusan duplikasi dan konflik label, pembagian data, ekstraksi fitur TF-IDF, pelatihan dan optimasi parameter Logistic Regression, serta evaluasi menggunakan data uji. Setiap tahapan memiliki fungsi yang saling berkaitan untuk menghasilkan model klasifikasi yang terukur dan dapat dianalisis secara sistematis.

Sumber dan Karakteristik Data

Data yang digunakan dalam penelitian ini berasal dari dataset Sephora Products and Skincare Reviews yang dipublikasikan oleh Nady Inky melalui platform Kaggle. Dataset tersebut berisi ulasan pengguna terhadap berbagai produk kecantikan dan perawatan kulit, termasuk informasi mengenai penilaian pengguna terhadap produk yang mereka gunakan. Berdasarkan naskah penelitian, data awal terdiri atas 285.412 entri yang dihimpun dari lima berkas ulasan. Dataset mencakup 541 produk, 100 merek, dan 13 atribut. Atribut tersebut meliputi rating, is_recommended, submission_time, review_text, review_title, skin_tone, eye_color, skin_type, hair_color, product_id, product_name, brand_name, dan price_usd.

Dari keseluruhan atribut tersebut, penelitian terutama menggunakan rating sebagai dasar penentuan label sentimen dan review_text sebagai sumber teks yang akan diproses oleh model. Atribut lainnya dapat digunakan untuk pemeriksaan karakteristik dataset, tetapi tidak otomatis menjadi fitur klasifikasi dalam eksperimen ini. Unit analisis penelitian adalah satu ulasan pengguna yang memiliki teks ulasan dan rating. Pemilihan unit analisis tersebut memungkinkan setiap sampel memiliki masukan berupa teks serta label yang digunakan dalam proses pelatihan dan evaluasi model.

Sebelum digunakan, dataset diperiksa untuk mengidentifikasi data yang tidak lengkap dan memastikan bahwa setiap sampel memenuhi kebutuhan penelitian. Sebanyak 345 baris yang tidak memiliki teks ulasan atau rating dikeluarkan dari data awal, sehingga tersisa 285.067 baris untuk diproses lebih lanjut. Setelah itu, dilakukan pemeriksaan duplikasi, teks yang menjadi kosong setelah prapemrosesan, dan konflik label. Berdasarkan hasil akhir yang dilaporkan dalam naskah, dataset yang digunakan dalam pemodelan berjumlah 245.092 ulasan.

Perlu diperiksa kembali bahwa selisih antara 285.067 baris setelah pemeriksaan awal dan 245.092 ulasan akhir mencapai 39.975 baris. Pengurangan tersebut perlu dijelaskan berdasarkan catatan pemrosesan aktual agar pembaca mengetahui berapa banyak data yang dihapus karena duplikasi, konflik label, teks kosong setelah pembersihan, atau alasan lainnya. Angka pengurangan tersebut tidak sebaiknya langsung dianggap seluruhnya sebagai data duplikat tanpa pemeriksaan lebih lanjut.

Pelabelan Sentimen

Pelabelan sentimen dilakukan menggunakan rating yang diberikan pengguna pada setiap produk. Pendekatan ini digunakan untuk membentuk kategori sentimen secara konsisten pada dataset berukuran besar tanpa harus memberikan anotasi manual pada seluruh ulasan. Penelitian ini menggunakan tiga kategori sentimen dengan aturan sebagai berikut.

Tabel 1. Pelabelan Sentimen

Label	Rating pengguna	Jumlah	Proporsi
Positif	4-5	203.896	83,19%
Netral	3	17.842	7,28%
Negatif	1–2	23.354	9,53%
Total		245.092	100%

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

Berdasarkan aturan tersebut, setiap ulasan yang memiliki rating 1 atau 2 dimasukkan ke kelas negatif. Ulasan dengan rating 3 dimasukkan ke kelas netral, sedangkan ulasan dengan rating 4 atau 5 dimasukkan ke kelas positif. Penggunaan rating sebagai dasar pelabelan memiliki keuntungan dari sisi konsistensi dan efisiensi. Dengan aturan yang telah ditetapkan, label dapat dibentuk secara otomatis untuk seluruh data yang memenuhi kriteria penelitian. Proses ini juga mengurangi kebutuhan anotasi manual dalam jumlah besar.

Namun, rating merupakan representasi penilaian pengguna terhadap produk dan tidak selalu mencerminkan keseluruhan sentimen yang terkandung dalam teks. Sebagai contoh, pengguna dapat memberikan rating tinggi tetapi tetap menyampaikan keluhan mengenai harga atau kemasan. Sebaliknya, pengguna dapat memberikan rating sedang meskipun sebagian besar isi ulasannya menunjukkan kepuasan.

Oleh sebab itu, label yang digunakan dalam penelitian ini perlu dipahami sebagai sentimen berbasis rating, bukan label sentimen yang telah diverifikasi melalui anotasi manual berdasarkan makna teks. Batasan tersebut penting karena model mempelajari hubungan antara teks ulasan dan kategori rating yang diturunkan menjadi label sentimen.

Pembersihan Data dan Prapemrosesan Teks

Prapemrosesan teks merupakan tahapan untuk mengubah ulasan mentah menjadi teks yang lebih konsisten sebelum dilakukan ekstraksi fitur. Ulasan pengguna umumnya memiliki variasi penulisan, penggunaan huruf kapital, URL, tanda baca, dan kosakata informal. Jika variasi tersebut tidak ditangani dengan aturan yang konsisten, representasi fitur dapat menjadi lebih besar atau memuat perbedaan yang tidak relevan bagi proses klasifikasi. Dalam penelitian ini, pembersihan data dilakukan melalui pemeriksaan kelengkapan data, penyeragaman teks, penghapusan elemen tertentu, penghapusan kata henti, serta pemeriksaan duplikasi dan konflik label.

Pemeriksaan data kosong

Tahap pertama adalah memeriksa kelengkapan kolom yang dibutuhkan untuk eksperimen, terutama review_text dan rating. Baris yang tidak memiliki teks ulasan atau rating dikeluarkan karena tidak dapat digunakan secara langsung untuk mempelajari hubungan antara teks dan label.

Berdasarkan hasil yang dilaporkan, sebanyak 345 baris dikeluarkan pada tahap pemeriksaan awal sehingga jumlah data berkurang dari 285.412 menjadi 285.067 baris. Pemeriksaan data kosong diperlukan karena model klasifikasi teks membutuhkan masukan yang dapat diubah menjadi fitur numerik. Sementara itu, rating diperlukan untuk menentukan kelas sentimen yang menjadi target prediksi.

Penyeragaman huruf (lowercasing)

Seluruh teks ulasan diubah menjadi huruf kecil untuk mengurangi perbedaan representasi kata pernyataan yang lebih hati-hati penggunaan huruf kapital. Sebagai contoh, kata Good, GOOD, dan good akan diseragamkan menjadi good. Tanpa proses ini, kata yang sebenarnya memiliki bentuk dan makna sama dapat diperlakukan sebagai fitur berbeda oleh representasi teks. Penyeragaman huruf membantu mengurangi variasi fitur yang tidak diperlukan dan membuat proses ekstraksi TF-IDF menjadi lebih konsisten.

Penghapusan URL

URL yang terdapat dalam teks ulasan dihapus karena penelitian berfokus pada isi linguistik ulasan, bukan alamat halaman web yang dicantumkan pengguna. Penghapusan URL membantu mengurangi fitur yang tidak secara langsung merepresentasikan penilaian pengguna terhadap produk. Akan tetapi, penerapan aturan ini perlu dilakukan secara konsisten agar bagian kata yang relevan tidak ikut terhapus.

Pembersihan karakter

Tahap selanjutnya adalah menghapus karakter yang tidak dipertahankan berdasarkan aturan pembersihan yang digunakan dalam eksperimen. Naskah menyebutkan bahwa karakter selain huruf dihapus sesuai aturan prapemrosesan yang diterapkan. Proses ini dapat membantu mengurangi variasi simbol yang tidak diperlukan dalam representasi teks. Namun, aturan pembersihan harus memperhatikan kemungkinan bahwa tanda baca tertentu dapat menunjukkan ekspresi atau penekanan sentimen.

Oleh karena itu, aturan pembersihan yang sebenarnya perlu mengikuti kode eksperimen, bukan hanya deskripsi umum. Jika angka, tanda baca, atau karakter tertentu dihapus, hal tersebut perlu dipastikan dari implementasi yang digunakan.

Penghapusan stopword

Stopword removal dilakukan untuk menghapus kata-kata umum yang dianggap kurang memberikan informasi pembeda dalam klasifikasi. Dalam penelitian ini, penghapusan stopword dilakukan menggunakan daftar kata henti bahasa Inggris karena ulasan pada dataset menggunakan bahasa Inggris.

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

Meskipun dapat mengurangi jumlah fitur, penghapusan stopword harus dilakukan dengan hati-hati. Beberapa kata yang sering muncul justru memiliki peran penting dalam menyampaikan sentimen, terutama kata negasi.

Mempertahankan kata negasi

Penelitian ini secara khusus mempertahankan kata negasi no, nor, dan not selama proses penghapusan stopword. Keputusan tersebut penting karena kata negasi dapat mengubah makna keseluruhan kalimat.

Tabel 2. Contoh Hasil Prapemrosesan Teks Ulasan

Teks Asli	Teks Setelah Prapemrosesan
Not worth the money, I would not recommend it.	not worth money would not recommend

Kata not tetap dipertahankan karena menunjukkan penilaian negatif. Jika kata tersebut dihapus, kalimat dapat berubah menjadi "worth money would recommend", sehingga makna teks yang tersisa berpotensi menunjukkan kecenderungan berbeda dari ulasan aslinya. Contoh tersebut memperlihatkan bahwa prapemrosesan tidak cukup hanya mengurangi jumlah kata. Proses ini juga harus mempertahankan unsur linguistik yang memiliki pengaruh terhadap makna sentimen.

Pemeriksaan teks kosong setelah prapemrosesan

Teks yang awalnya memiliki isi dapat menjadi kosong setelah URL, karakter tertentu, atau kata henti dihapus. Oleh sebab itu, pemeriksaan ulang dilakukan setelah pembersihan teks. Baris yang tidak lagi memiliki teks yang dapat digunakan tidak dapat memberikan fitur linguistik bagi model. Dalam naskah, jumlah data berkurang dari 285.067 menjadi 245.092 setelah tahap ini, sehingga terdapat empat baris yang perlu dijelaskan sebagai teks kosong pascaprapemrosesan.

Penanganan duplikasi

Pemeriksaan duplikasi dilakukan untuk mengidentifikasi ulasan yang sama atau berulang dalam dataset. Data duplikat dapat memengaruhi distribusi sampel dan berpotensi membuat evaluasi terlalu optimistis jika ulasan yang sama muncul dalam data latih dan data uji. Naskah melaporkan adanya 37.611 kelompok teks ulasan yang berulang dengan total 40.157 baris duplikat. Sebanyak 39.959 baris duplikat dilaporkan dihilangkan. Karena jumlah baris yang disebut berulang dan jumlah yang dihapus tidak identik, definisi kelompok duplikasi dan prosedur penghapusannya perlu diperiksa kembali pada kode atau log eksperimen.

Secara metodologis, penghapusan duplikasi perlu menerapkan aturan yang jelas, misalnya berdasarkan teks ulasan yang telah dibersihkan. Jika beberapa ulasan identik memiliki rating berbeda, data tersebut perlu ditangani secara terpisah karena mungkin menunjukkan konflik label.

Penanganan konflik label

Konflik label terjadi ketika teks ulasan yang identik memiliki label sentimen berbeda. Kondisi tersebut dapat menyebabkan model menerima masukan teks yang sama dengan target prediksi berbeda. Naskah melaporkan lima kelompok data yang mengalami konflik label, terdiri atas 12 baris. Seluruh baris yang berkonflik tersebut dilaporkan dihapus untuk menjaga konsistensi hubungan antara teks dan label.

Keputusan ini membantu mengurangi ambiguitas dalam data pelatihan. Namun, penghapusan konflik label sebaiknya didasarkan pada aturan yang ditentukan sebelum pemodelan dan dicatat secara transparan agar prosesnya dapat direproduksi.

Pembagian Dataset

Setelah tahapan pembersihan data selesai, dataset akhir berjumlah 245.092 ulasan dibagi menjadi data latih dan data uji dengan rasio 80:20. Pembagian ini bertujuan memisahkan data yang digunakan untuk membangun model dari data yang digunakan untuk mengukur kemampuan prediksi model pada sampel yang tidak digunakan dalam pelatihan.

Data latih berjumlah 196.073 ulasan, sedangkan data uji berjumlah 49.019 ulasan. Pembagian dilakukan dengan metode stratified split, yaitu mempertahankan proporsi masing-masing kelas sentimen agar distribusi kelas dalam data latih dan data uji tetap mendekati distribusi dataset awal. Rincian pembagian data berdasarkan naskah adalah sebagai berikut.

Tabel 3. Distribusi Data Sentimen pada Data Latih dan Data Uji

Kelas sentimen	Data latih	Data uji	Total
Positif	163.116	40.780	203.896
Netral	14.274	3.568	17.842
Negatif	18.683	4.671	23.354
Total	196.073	49.019	245.092

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

Pembagian secara stratifikasi penting karena distribusi kelas pada dataset tidak seimbang. Jika pembagian dilakukan tanpa mempertimbangkan label, jumlah kelas netral atau negatif pada data uji dapat menjadi terlalu sedikit sehingga hasil evaluasi kurang representatif. Data latih digunakan untuk mempelajari hubungan antara fitur teks dan label sentimen. Data uji dipertahankan untuk evaluasi akhir dan tidak boleh digunakan untuk memilih parameter terbaik.

Untuk menghindari data leakage, seluruh parameter ekstraksi TF-IDF harus dipelajari dari data latih. Setelah proses pembelajaran selesai, transformasi yang sama diterapkan pada data uji. Dengan demikian, informasi statistik dari data uji tidak ikut memengaruhi pembentukan fitur pada tahap pelatihan. Naskah menyebutkan penggunaan stratifikasi, tetapi nilai random_state tidak tercantum dalam bagian metode yang tersedia. Faktual berdasarkan kode dan log eksperimen, nilainya perlu dicantumkan agar pembagian data dapat direproduksi.

Ekstraksi Fitur Menggunakan TF-IDF

Setelah teks dibersihkan dan dataset dibagi, ulasan diubah menjadi representasi numerik menggunakan Term Frequency–Inverse Document Frequency (TF-IDF). Tahapan ini diperlukan karena algoritma Logistic Regression tidak dapat memproses teks mentah secara langsung dan membutuhkan fitur numerik sebagai masukan.

TF-IDF memberikan bobot pada istilah berdasarkan frekuensi kemunculannya dalam suatu dokumen dan tingkat penyebarannya dalam keseluruhan korpus. Kata yang sering muncul dalam satu ulasan tetapi relatif lebih jarang ditemukan pada dokumen lain dapat memperoleh bobot lebih besar daripada kata yang muncul hampir di seluruh dokumen. Secara umum, nilai TF-IDF dapat dinyatakan dengan rumus:

$$\mathrm{TFIDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$$

Keterangan:
t = istilah atau kata yang dievaluasi.
d = dokumen atau ulasan.
$\mathrm{TF}(t, d)$ = frekuensi istilah dalam dokumen.
$\mathrm{IDF}(t)$ = tingkat kelangkaan istilah dalam kumpulan dokumen.

Istilah yang sering muncul dalam suatu ulasan dapat memperoleh bobot TF lebih tinggi. Namun, jika istilah tersebut juga muncul pada hampir seluruh dokumen, nilai IDF-nya cenderung lebih rendah. Sebaliknya, istilah yang relatif jarang muncul pada keseluruhan korpus dapat memiliki bobot lebih besar apabila cukup sering muncul pada dokumen tertentu. Berdasarkan naskah penelitian, konfigurasi TF-IDF yang digunakan mencakup parameter berikut.

Tabel 4. Parameter dan Konfigurasi TF-IDF dalam Ekstraksi Fitur Teks.

Parameter	Nilai	Fungsi
max_features	20.000	Membatasi jumlah fitur yang dipertahankan paling banyak 20.000 fitur.
ngram_range	(1, 2)	Mempertimbangkan unigram dan bigram.
min_df	3	Mempertahankan istilah yang muncul pada sekurang-kurangnya tiga dokumen.

Penggunaan unigram memungkinkan model mempelajari kata tunggal seperti good, dry, atau irritation. Sementara itu, bigram memungkinkan model mengenali kombinasi dua kata, seperti not good atau very dry, yang dapat memberikan konteks lebih spesifik daripada kata tunggal. Parameter min_df=3 digunakan untuk mengurangi fitur yang terlalu jarang muncul. Parameter max_features=20000 membatasi dimensi representasi teks agar jumlah fitur tetap terkendali. Pembatasan ini dapat membantu mengelola kompleksitas komputasi, meskipun sebagian istilah yang jarang muncul tetapi informatif juga mungkin tidak dipertahankan.

Dalam penerapannya, TF-IDF perlu di-fit hanya pada data latih. Data uji kemudian ditransformasikan menggunakan kosakata dan bobot IDF yang telah dipelajari dari data latih. Jika TF-IDF dipelajari menggunakan keseluruhan dataset sebelum pembagian data, proses tersebut dapat menyebabkan kebocoran informasi dan membuat evaluasi kurang mencerminkan kondisi penggunaan model pada data baru. Naskah juga melaporkan bahwa 20.000 fitur yang terbentuk terdiri atas 4.473 unigram dan 15.527 bigram. Angka tersebut dapat dipertahankan selama sesuai dengan keluaran aktual dari proses ekstraksi fitur.

Klasifikasi Menggunakan Logistic Regression

Logistic Regression digunakan sebagai algoritma utama untuk mengklasifikasikan ulasan ke dalam tiga kategori sentimen. Meskipun namanya mengandung istilah regresi, algoritma ini dapat digunakan untuk tugas klasifikasi karena menghasilkan probabilitas keanggotaan suatu sampel pada kelas tertentu. Pada penelitian ini, masukan model berupa vektor fitur TF-IDF dari teks ulasan, sedangkan target prediksinya adalah label

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

positif, netral, atau negatif. Selama pelatihan, model mempelajari bobot yang menghubungkan fitur kata dengan kelas sentimen berdasarkan data latih.

Dalam klasifikasi multikelas, Logistic Regression dapat memperkirakan probabilitas untuk setiap kelas. Prediksi akhir umumnya ditentukan berdasarkan kelas dengan probabilitas tertinggi. Dengan demikian, model akan memilih salah satu kelas sentimen berdasarkan pola fitur yang dipelajari selama pelatihan. Salah satu kelebihan Logistic Regression adalah kemampuannya mengolah fitur teks berdimensi tinggi dan menghasilkan model yang relatif efisien. Kombinasinya dengan TF-IDF juga memungkinkan model mempelajari hubungan antara kata-kata dalam ulasan dan label sentimen tanpa memerlukan representasi teks berbasis jaringan saraf yang kompleks.

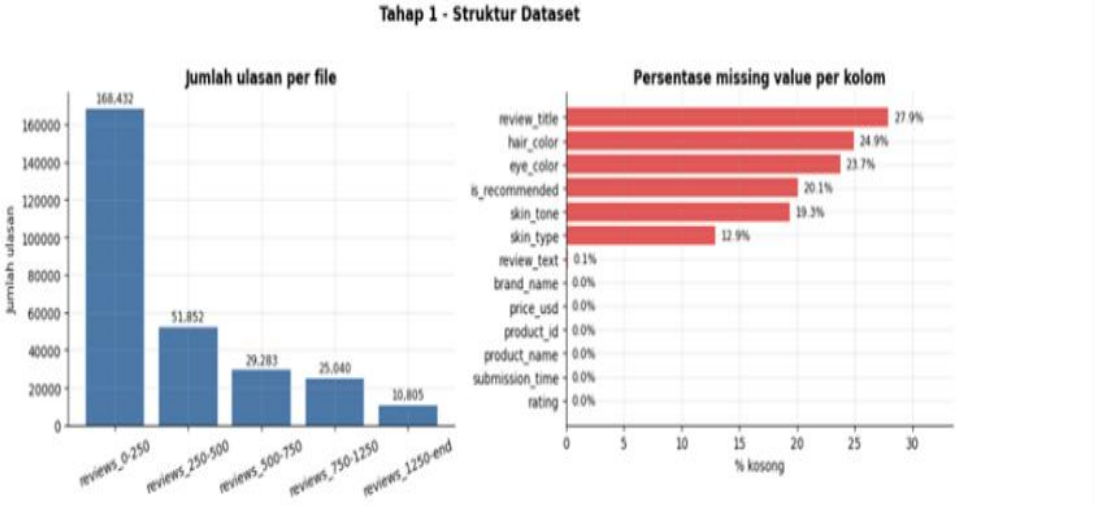
Namun, kemampuan model tetap dipengaruhi oleh kualitas data, distribusi kelas, pemilihan fitur, serta karakteristik bahasa dalam ulasan. Kata yang sama dapat digunakan dalam konteks berbeda, sedangkan ulasan netral dapat memuat campuran ungkapan positif dan negatif. Karena itu, model perlu dievaluasi berdasarkan metrik per kelas, bukan hanya berdasarkan akurasi keseluruhan.

Optimasi Parameter Menggunakan GridSearchCV

Pencarian parameter dilakukan menggunakan GridSearchCV dengan Stratified 3-Fold Cross-Validation pada 60.000 sampel yang dipilih secara acak dari data latih. Parameter yang diuji meliputi nilai C sebesar 0,1; 0,3; 1; 3; dan 10, serta dua pilihan pembobotan kelas, yaitu tanpa pembobotan dan `balanced`. Kinerja setiap konfigurasi dinilai menggunakan macro-F1. Hasil pencarian menunjukkan bahwa konfigurasi terbaik adalah C = 3,0 dengan `class_weight='balanced'`, yang memperoleh rata-rata macro-F1 sebesar 0,6776 dengan standar deviasi 0,0007. Setelah parameter terbaik ditentukan, model akhir dilatih menggunakan seluruh data latih.

HASIL DAN PEMBAHASAN

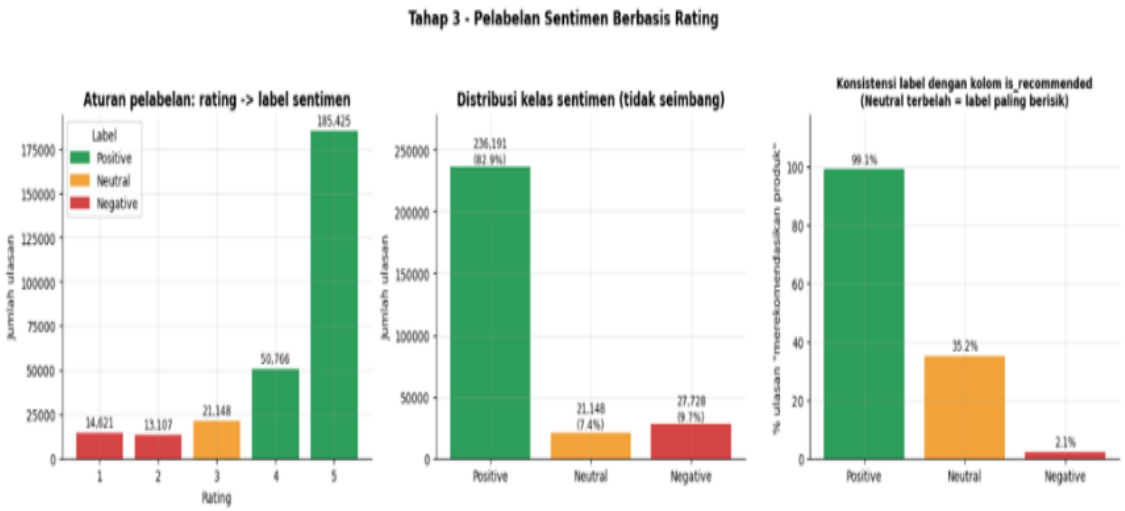
Data yang digunakan dalam penelitian ini berupa ulasan produk kecantikan yang dikumpulkan dari lima berkas data mentah dengan total keseluruhan sebanyak 285.412 entri. Dataset ini mencakup 13 atribut utama, yaitu rating, is_recommended, submission_time, review_text, review_title, skin_tone, eye_color, skin_type, hair_color, product_id, product_name, brand_name, dan price_usd. Pada distribusi rating awal, ulasan didominasi oleh pemberian rating 5 sebanyak 185.671 ulasan, diikuti oleh rating 4 sebanyak 50.814 ulasan, rating 3 sebanyak 21.164 ulasan, rating 1 sebanyak 14.645 ulasan, dan rating 2 sebanyak 13.118 ulasan. Gambaran mengenai rincian jumlah ulasan tiap berkas beserta tingkat kelengkapan data disajikan pada Gambar 1.



Gambar 1. Grafik Jumlah Ulasan per Berkas & Persentase Missing Value per Kolom

Berdasarkan pengelompokan rating ke dalam label sentimen awal, sampel ulasan memperlihatkan ketidakseimbangan kelas yang cukup signifikan. Sentimen positif yang diturunkan dari rating 4 dan 5 mencakup porsi terbesar yaitu sebanyak 236.191 ulasan (82,85%). Sementara itu, sentimen negatif yang berasal dari rating 1 dan 2 berjumlah 27.728 ulasan (9,73%), dan sentimen netral dari rating 3 berjumlah 21.148 ulasan (7,42%). Visualisasi mengenai proporsi distribusi rating dan sentimen awal ini diperlihatkan pada Gambar 2. Sebelum melangkah ke tahap prapemrosesan teks, dilakukan pemeriksaan kelengkapan data teks yang mengidentifikasi adanya 345 baris ulasan tanpa isi (missing review_text). Baris-baris tersebut langsung dihapus dari dataset sehingga menyisakan 285.067 ulasan untuk diproses lebih lanjut.

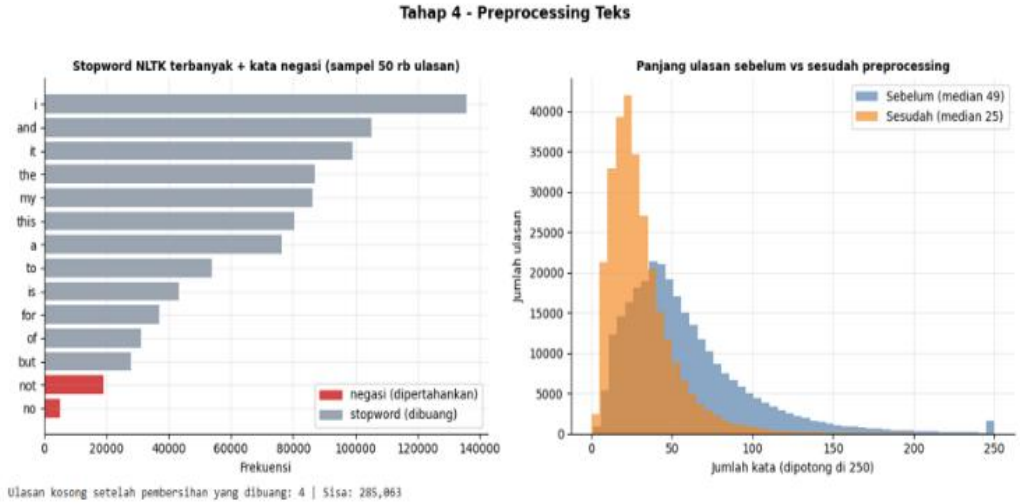
e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498



Gambar 2. Chart Distribusi Rating dan Proporsi Sentimen Awal

Hasil Prapemrosesan Teks dan Pembersihan Data

Tahap prapemrosesan teks dilakukan untuk membersihkan dan menstandarkan isi ulasan agar dapat diterima secara optimal oleh algoritma pembelajaran mesin. Sebagai contoh, kalimat ulasan mentah seperti "Not worth the money, I would not recommend it" ditransformasikan menjadi bentuk terstandardisasi "not worth money would not recommend". Dalam eksperimen ini, kata-kata negasi seperti not dan no secara sengaja dipertahankan dan tidak dibuang pada proses pembersihan kata henti (stopword removal). Kebijakan ini terbukti sangat krusial dalam menjaga polaritas makna kalimat. Jika kata negasi dibuang, teks ulasan tersebut akan berubah menjadi "worth money would recommend" yang mengubah total konteks kalimat dari sentimen negatif menjadi positif. Perbandingan perubahan panjang teks ulasan sebelum dan sesudah prapemrosesan disajikan pada Gambar 3.

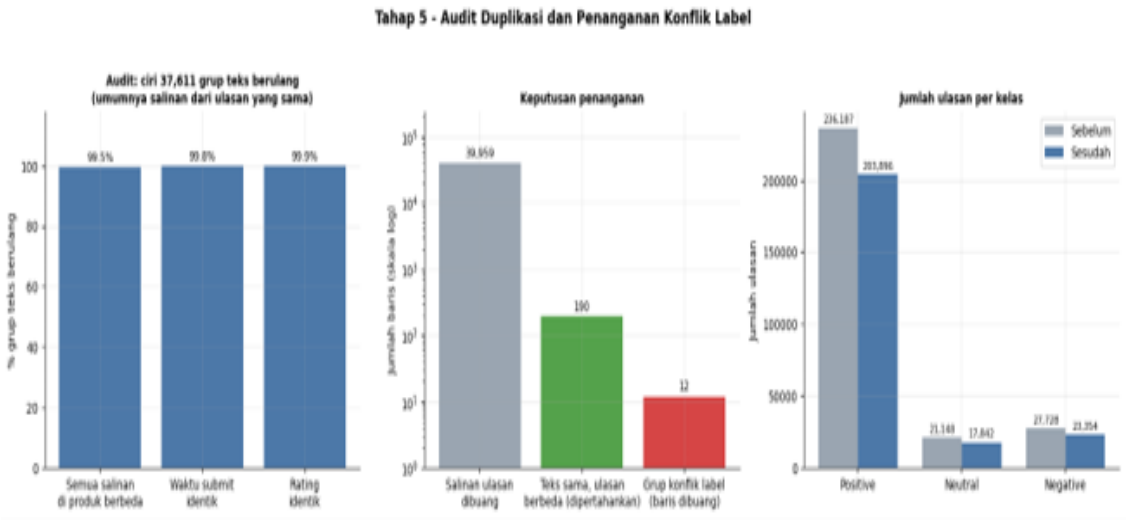


Gambar 3. Grafik Perbandingan Panjang Ulasan Sebelum dan Sesudah Preprocessing

Setelah tahap pembersihan teks, analisis duplikasi data menunjukkan adanya 37.611 kelompok teks ulasan yang berulang dengan total 40.157 baris duplikat. Dari temuan ini, sebanyak 39.959 baris ulasan duplikat dihilangkan. Selain itu, terdeteksi pula 5 kelompok data (terdiri dari 12 baris) yang mengalami konflik label, di mana teks ulasan yang persis sama memiliki label sentimen yang berbeda. Seluruh baris yang berkonflik tersebut dihapus demi menjaga konsistensi pelabelan.

Secara rinci, alur perubahan jumlah data berlangsung dari 285.412 ulasan awal, berkurang menjadi 285.067 ulasan setelah eliminasi teks kosong, lalu menjadi 245.092 ulasan setelah penghapusan teks yang menjadi kosong pasca-prapemrosesan, hingga akhirnya menghasilkan dataset bersih sebanyak 245.092 ulasan. Dataset akhir ini terdiri dari 203.896 ulasan berlabel positif, 17.842 ulasan berlabel netral, dan 23.354 ulasan berlabel negatif. Alur pergeseran jumlah data di setiap tahapannya diilustrasikan pada Gambar 4.

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498



Gambar 4. Diagram Alir / Bar Chart Perubahan Jumlah Data Tiap Tahap Pemrosesan

Hasil Pelatihan dan Pengujian Model

Dataset bersih sebanyak 245.092 ulasan selanjutnya dibagi menjadi data latih dan data uji menggunakan metode Stratified Split dengan rasio pembagian 80% : 20%. Pembagian ini menghasilkan 196.073 data latih terdiri atas 163.116 ulasan positif, 14.274 ulasan netral, dan 18.683 ulasan negatif serta 49.019 data uji terdiri atas 40.780 ulasan positif, 3.568 ulasan netral, dan 4.671 ulasan negatif.

Proses pencarian hyperparameter terbaik (hyperparameter tuning) dilakukan pada data latih menggunakan metode 3-Fold Cross-Validation. Kombinasi parameter terbaik yang diperoleh dari hasil eksplorasi adalah nilai regularisasi C = 3.0 dan penerapan penyeimbangan kelas class_weight = 'balanced'. Kombinasi ini memberikan performa rata-rata skor Macro-F1 sebesar 0,6776 (67,76%) pada proses validasi silang dengan tingkat variasi standar deviasi yang sangat kecil yaitu 0,0007. Model akhir kemudian dibangun menggunakan pembentukan fitur TF-IDF berkapasitas 20.000 fitur yang mencakup 4.473 unigram dan 15.527 bigram serta dikombinasikan dengan konfigurasi model Logistic Regression terpilih.

Evaluasi Kinerja Model

Pengujian model Logistic Regression pada 49.019 data uji menghasilkan performa klasifikasi yang bervariasi pada setiap kelas sentimen, sebagaimana dirangkum pada Tabel 5.

Tabel 5. Hasil Evaluasi Performa Model pada Data Uji

Kelas Sentimen	Precision	Recall	F1-Score
Positif	0,977	0,901	0,938
Netral	0,325	0,552	0,409
Negatif	0,642	0,738	0,687
Macro Average	0,648	0,730	0,678
Weighted Average	0,898	0,860	0,875

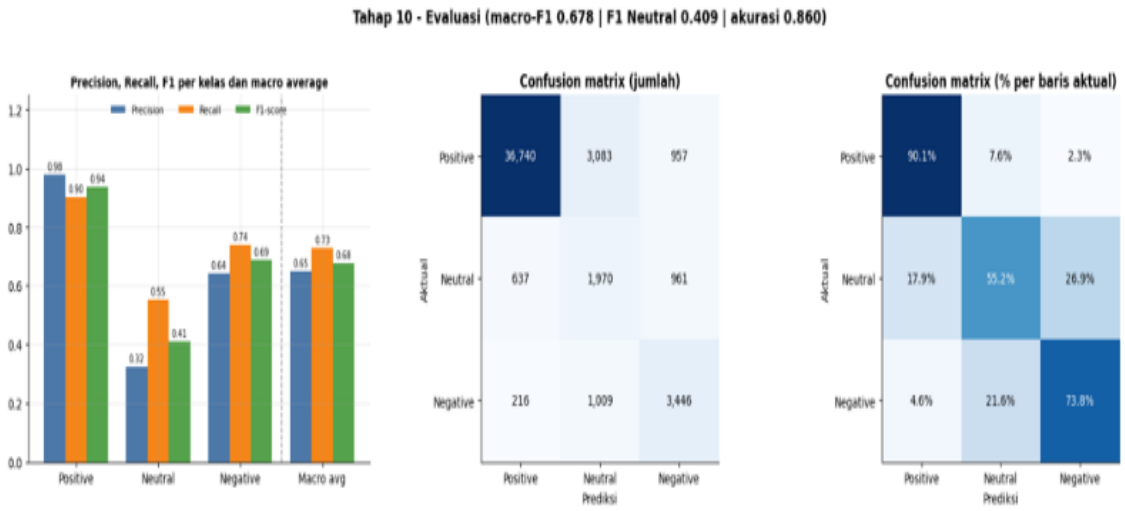
Berdasarkan Tabel 5, model Logistic Regression menunjukkan hasil evaluasi yang berbeda pada setiap kelas sentimen. Kelas positif memperoleh F1-score tertinggi sebesar 0,938, diikuti kelas negatif sebesar 0,687, sedangkan kelas netral memperoleh F1-score terendah sebesar 0,409. Hasil ini menunjukkan bahwa model lebih baik dalam mengklasifikasikan ulasan positif dibandingkan ulasan netral.

Secara keseluruhan, model memperoleh akurasi sebesar 86,00% dan weighted-F1 sebesar 0,875, tetapi nilai macro-F1 hanya mencapai 0,678. Perbedaan ini menunjukkan bahwa nilai evaluasi keseluruhan belum sepenuhnya menggambarkan kemampuan model pada setiap kelas. Weighted-F1 mempertimbangkan jumlah sampel pada masing-masing kelas, sedangkan macro-F1 memberikan bobot yang sama kepada setiap kelas. Karena data didominasi oleh ulasan positif, nilai weighted-F1 cenderung lebih mencerminkan kinerja pada kelas tersebut. Sementara itu, macro-F1 memperlihatkan bahwa kinerja model belum merata pada ketiga kelas sentimen.

Rendahnya F1-score kelas netral menunjukkan bahwa pengenalan ulasan netral masih menjadi tantangan bagi model. Ketidakseimbangan jumlah data mungkin berkaitan dengan perbedaan kinerja tersebut, tetapi hasil evaluasi ini belum cukup untuk memastikan bahwa ketidakseimbangan data merupakan penyebab utamanya. Selain itu, label sentimen penelitian dibentuk berdasarkan rating pengguna, sehingga kategori yang diprediksi

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

model belum tentu sepenuhnya sesuai dengan sentimen yang benar-benar dinyatakan dalam teks ulasan. Oleh karena itu, hasil akurasi 86,00% perlu dibaca bersama macro-F1 dan nilai evaluasi per kelas agar keterbatasan model dapat terlihat dengan lebih jelas. Dapat dilihat secara visual pada Gambar 5.



Gambar 5. Heatmap Confusion Matrix Hasil Pengujian Model Logistic Regression

Performa klasifikasi pada kelas netral masih lebih rendah (F1-score 0,409) dibandingkan dengan kelas positif (0,938) dan kelas negatif (0,687). Salah satu karakteristik dataset yang perlu dipertimbangkan adalah ketidakseimbangan distribusi kelas, karena ulasan netral hanya mencakup sekitar 7,3% dari data uji. Selain itu, kemungkinan adanya variasi dan ambiguitas bahasa dalam ulasan netral juga perlu dikaji lebih lanjut. Namun, hasil evaluasi ini belum cukup untuk memastikan bahwa kedua faktor tersebut merupakan penyebab langsung rendahnya kinerja model pada kelas netral. Pemeriksaan terhadap sampel yang salah diklasifikasikan diperlukan untuk memperoleh pemahaman yang lebih mendalam.

KESIMPULAN

Penelitian ini berhasil menerapkan algoritma Logistic Regression yang dikombinasikan dengan ekstraksi fitur TF-IDF untuk mengklasifikasikan ulasan pengguna produk skincare Sephora ke dalam tiga kelas sentimen, yaitu positif, netral, dan negatif. Berdasarkan hasil evaluasi pada data uji, model mampu mencapai kinerja utama berupa tingkat akurasi sebesar 86,00%, skor macro-F1 sebesar 0,6778, dan weighted-F1 sebesar 0,8752. Hasil penelitian menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengenali sentimen positif (F1-score 0,938), namun mengalami keterbatasan dalam mengidentifikasi sentimen netral (F1-score 0,409) pernyataan yang lebih hati-hati adanya ketidakseimbangan distribusi data serta ambiguitas bahasa dalam ulasan. Temuan ini berhasil menjawab tujuan penelitian mengenai efektivitas dan karakteristik kinerja Logistic Regression pada klasifikasi sentimen berbasis rating. Perlu ditegaskan bahwa hasil evaluasi ini bersifat spesifik pada karakteristik dan cakupan dataset ulasan Sephora yang digunakan, sehingga tidak dimaksudkan untuk digeneralisasi pada dataset atau platform lain di luar batas penelitian ini.

Untuk penelitian selanjutnya, disarankan untuk melakukan validasi label sentimen melalui anotasi manual pada sebagian data agar kesesuaian antara rating dan makna ulasan dapat dievaluasi. Penelitian berikutnya juga dapat menguji strategi penanganan ketidakseimbangan kelas serta menyesuaikan proses prapemrosesan untuk mempertahankan informasi linguistik yang penting, terutama pada ulasan netral. Selain itu, pengujian pada pembagian data yang lebih ketat berdasarkan produk atau waktu ulasan dapat dilakukan untuk menilai kemampuan generalisasi model terhadap data yang berbeda. Langkah-langkah tersebut diharapkan dapat membantu meningkatkan keandalan evaluasi dan pemahaman terhadap keterbatasan klasifikasi sentimen menggunakan Logistic Regression.

REFERENSI

Fadly, D., Marlina, Kurniawan, T. B., Zakaria, M. Z., Binti, S. F., & Abdullah. (2022). Pandangan tentang analisis sentimen terhadap produk perawatan kulit alami. Journal of Data Science, 1(12), 1–17.

Dikusuma, & Widjaja. (2025). Perbandingan Naïve Bayes dan regresi logistik untuk analisis sentimen review produk skincare di Tokopedia. Jurnal Informatika dan Teknologi Komputer, 5(3), 1–10.

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study.

e-ISSN: 2963 – 198X
p-ISSN: 2963 – 2498

Artificial Intelligence Review, 54(7). https://doi.org/10.1007/s10462-021-09973-3

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences, 13(7). https://doi.org/10.3390/app13074550