



Integrasi Principal Component Analysis dan Logistic Regression untuk Analisis Sentimen Kepuasan Pelanggan Berdasarkan Ulasan Online

Tsalsabila Jilhan Haura*, Rini Sovia, Gunadi Widi Nurcahyo

Fakultas Ilmu Komputer, Magister Teknik Informatika, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

Email: ^{1,*}jilhanhaura07@gmail.com, ²rini_sovia@upiypk.ac.id, ³gunadiwidi@yahoo.co.id

Email Penulis Korespondensi: jilhanhaura07@gmail.com

Abstrak—Ulasan pelanggan pada platform digital merupakan sumber informasi yang penting untuk mengevaluasi kualitas layanan dan tingkat kepuasan pelanggan. Namun, karakteristik data ulasan yang tidak terstruktur dan memiliki dimensi fitur yang tinggi menjadi tantangan dalam proses analisis sentimen. Penelitian ini bertujuan mengembangkan model analisis sentimen pelanggan dengan mengintegrasikan Principal Component Analysis (PCA) dan Logistic Regression. Data yang digunakan berupa 679 ulasan berbahasa Indonesia yang diperoleh melalui teknik web scraping dari Google Reviews pada sepuluh cabang d'Besto EBM di Kota Padang. Tahapan penelitian meliputi preprocessing teks, pembobotan TF-IDF, reduksi dimensi menggunakan PCA, serta klasifikasi sentimen menggunakan Logistic Regression. Hasil penelitian menunjukkan bahwa PCA mampu mereduksi kompleksitas data dengan menghasilkan dua komponen utama yang menjelaskan 85,7% total variansi data. Model Logistic Regression yang dibangun pada fitur hasil reduksi PCA memperoleh akurasi sebesar 82%, yang menunjukkan kemampuan model dalam mengklasifikasikan sentimen positif dan negatif secara baik. Selain meningkatkan efisiensi komputasi, penggunaan PCA juga membantu mengurangi redundansi fitur pada data teks berdimensi tinggi. Kontribusi penelitian ini adalah menghasilkan pendekatan analisis sentimen yang lebih sederhana dan efisien untuk mengolah ulasan pelanggan serta menyediakan informasi berbasis data yang dapat dimanfaatkan sebagai pendukung evaluasi kualitas layanan dan pengambilan keputusan pada industri kuliner.

Kata Kunci: Analisis Sentimen; Logistic Regression; Principal Component Analysis; d'Besto EBM; Google Reviews

Abstract—Customer reviews on digital platforms are an important source of information for evaluating service quality and customer satisfaction levels. However, the unstructured nature of review data and its high feature dimensionality pose challenges in the sentiment analysis process. This study aims to develop a customer sentiment analysis model by integrating Principal Component Analysis (PCA) and Logistic Regression. The data used are 679 Indonesian-language reviews obtained through web scraping techniques from Google Reviews at ten d'Besto EBM branches in Padang City. The research stages include text preprocessing, TF-IDF weighting, dimensionality reduction using PCA, and sentiment classification using Logistic Regression. The results show that PCA is able to reduce data complexity by producing two principal components that explain 85.7% of the total data variance. The Logistic Regression model built on the features resulting from PCA reduction achieved an accuracy of 82%, demonstrating the model's ability to effectively classify positive and negative sentiments. In addition to improving computational efficiency, the use of PCA also helps reduce feature redundancy in high-dimensional text data. The contribution of this research is to produce a simpler and more efficient sentiment analysis approach to process customer reviews and provide data-based information that can be used to support service quality evaluation and decision-making in the culinary industry.

Keywords: Sentiment Analysis; Logistic Regression; Principal Component Analysis; d'Besto EBM; Google Reviews

1. PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi telah mendorong transformasi digital pada berbagai sektor industri, termasuk industri kuliner. Pemanfaatan platform digital tidak hanya digunakan sebagai media promosi dan transaksi, tetapi juga menjadi sarana bagi pelanggan untuk memberikan ulasan terhadap produk maupun layanan yang mereka terima [1]. Ulasan pelanggan yang dipublikasikan melalui berbagai platform daring seperti Google Reviews, media sosial, dan aplikasi pemesanan makanan mengandung informasi penting yang mencerminkan persepsi, pengalaman, serta tingkat kepuasan pelanggan terhadap suatu bisnis [2]. Informasi tersebut dapat menjadi sumber umpan balik yang berharga bagi perusahaan dalam melakukan evaluasi dan perbaikan layanan secara berkelanjutan [3]. Meningkatnya jumlah pengguna internet di Indonesia turut mendorong peningkatan aktivitas masyarakat dalam memberikan ulasan secara daring. Berdasarkan data Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia telah mencapai lebih dari 210 juta jiwa atau sekitar 77,02% dari total populasi [4]. Kondisi ini menghasilkan volume data ulasan yang terus bertambah dan berpotensi dimanfaatkan sebagai sumber informasi strategis bagi organisasi. Namun, karakteristik data ulasan yang berbentuk teks tidak terstruktur menyebabkan proses analisis secara manual menjadi kurang efektif, terutama ketika jumlah data yang dianalisis semakin besar [5]. Selain membutuhkan waktu yang lama, pendekatan manual juga rentan terhadap subjektivitas sehingga sulit menghasilkan informasi yang konsisten dan terukur [6].

Salah satu pendekatan yang banyak digunakan untuk mengolah data ulasan pelanggan adalah analisis sentimen. Analisis sentimen merupakan teknik yang digunakan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini pengguna ke dalam kategori sentimen tertentu, seperti positif dan negatif [7]. Melalui analisis sentimen, perusahaan dapat memperoleh gambaran umum mengenai persepsi pelanggan terhadap kualitas produk maupun layanan yang diberikan [8]. Namun demikian, proses analisis sentimen pada data teks sering menghadapi tantangan berupa tingginya jumlah fitur yang dihasilkan dari representasi teks, khususnya ketika menggunakan teknik pembobotan seperti Term Frequency-Inverse Document Frequency (TF-IDF) [9]. Dimensi fitur yang sangat besar dapat meningkatkan kompleksitas komputasi, memperbesar risiko overfitting, dan menurunkan efisiensi model klasifikasi [10].

Diperlukan teknik reduksi dimensi yang mampu menyederhanakan representasi data tanpa menghilangkan informasi penting yang terkandung di dalamnya. Salah satu metode reduksi dimensi yang banyak digunakan adalah

Principal Component Analysis (PCA). PCA merupakan teknik statistik yang mentransformasikan sekumpulan variabel yang saling berkorelasi menjadi sejumlah komponen utama baru yang tidak berkorelasi dan mampu mempertahankan sebagian besar informasi dari data asli [11]. Dalam pengolahan data teks, PCA dapat digunakan untuk mengurangi jumlah fitur sehingga proses pembelajaran model menjadi lebih efisien tanpa kehilangan karakteristik utama data [12].

Setelah proses reduksi dimensi dilakukan, diperlukan metode klasifikasi yang mampu mengidentifikasi kategori sentimen secara efektif. Logistic Regression merupakan salah satu algoritma klasifikasi yang banyak digunakan karena memiliki proses komputasi yang relatif sederhana, mudah diimplementasikan, serta menghasilkan model yang mudah diinterpretasikan [13]. Logistic Regression bekerja dengan memodelkan probabilitas suatu data masuk ke dalam kelas tertentu berdasarkan hubungan antara variabel independen dan variabel dependen [14]. Pada kasus analisis sentimen, metode ini dinilai sesuai untuk mengklasifikasikan opini pelanggan ke dalam kategori sentimen positif maupun negatif [15]. Integrasi PCA dan Logistic Regression diharapkan mampu menghasilkan model yang tidak hanya memiliki performa klasifikasi yang baik, tetapi juga lebih efisien dalam menangani data teks berdimensi tinggi [16].

Penelitian oleh [17] menunjukkan bahwa penerapan PCA sebelum proses klasifikasi mampu meningkatkan efisiensi pengolahan data teks dan mengurangi kompleksitas fitur tanpa menurunkan performa model secara signifikan. Penelitian lain melaporkan bahwa penggunaan feature selection berbasis Particle Swarm Optimization (PSO) pada Support Vector Machine (SVM) dapat meningkatkan akurasi klasifikasi dari 82,51% menjadi 86,07% [18]. Selain itu, pendekatan hibrida Convolutional Neural Network dan Support Vector Machine (CNN-SVM) mampu mencapai akurasi sebesar 98,32% pada tugas klasifikasi tertentu [19]. Meskipun demikian, sebagian besar penelitian terdahulu berfokus pada peningkatan akurasi model dengan menggunakan algoritma yang lebih kompleks, sementara kajian mengenai efektivitas kombinasi metode reduksi dimensi dan algoritma klasifikasi yang lebih sederhana masih relatif terbatas [20].

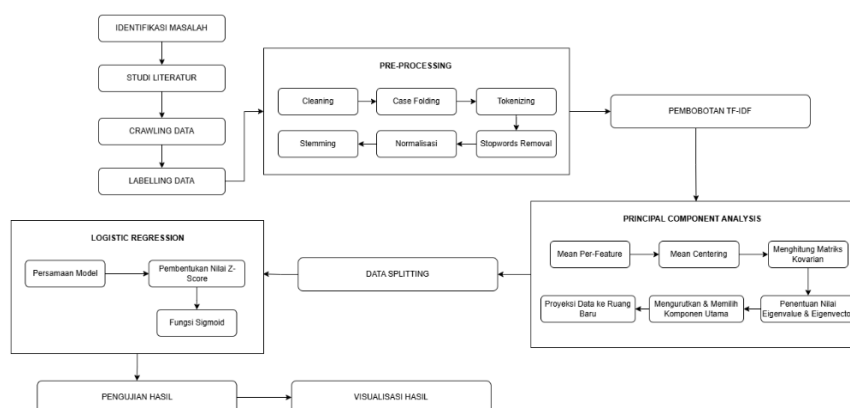
Berdasarkan kajian literatur tersebut, terdapat celah penelitian yang menunjukkan bahwa hubungan antara reduksi dimensi menggunakan PCA dan performa klasifikasi Logistic Regression pada data ulasan pelanggan masih belum banyak dieksplorasi, khususnya pada data ulasan berbahasa Indonesia. Sebagian besar penelitian lebih menitikberatkan pada perbandingan algoritma klasifikasi atau penggunaan metode deep learning yang membutuhkan sumber daya komputasi lebih besar. Padahal, pendekatan yang lebih sederhana dan mudah diinterpretasikan tetap memiliki nilai penting, terutama bagi organisasi yang membutuhkan solusi analitik yang efisien dan mudah diterapkan [21]. Oleh karena itu, diperlukan penelitian yang mengevaluasi kemampuan PCA dalam menyederhanakan representasi fitur teks sekaligus mengukur performa Logistic Regression dalam melakukan klasifikasi sentimen pada data ulasan pelanggan.

Penelitian ini bertujuan mengembangkan model analisis sentimen pelanggan dengan mengintegrasikan Principal Component Analysis (PCA) sebagai metode reduksi dimensi dan Logistic Regression sebagai metode klasifikasi. Data penelitian diperoleh dari ulasan pelanggan yang dikumpulkan melalui platform Google Reviews. Kontribusi penelitian ini terletak pada evaluasi efektivitas PCA dalam mengurangi kompleksitas data teks serta pengukuran kinerja Logistic Regression pada fitur hasil reduksi dimensi. Hasil penelitian diharapkan dapat memberikan alternatif pendekatan yang efisien dan mudah diinterpretasikan untuk analisis sentimen pada data ulasan pelanggan, serta menjadi dasar dalam pemanfaatan data ulasan sebagai pendukung pengambilan keputusan berbasis data.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan text mining untuk melakukan analisis sentimen terhadap ulasan pelanggan yang diperoleh dari platform Google Reviews. Metode yang digunakan adalah Principal Component Analysis (PCA) sebagai teknik reduksi dimensi dan Logistic Regression sebagai algoritma klasifikasi sentimen. Tahapan penelitian dimulai dari identifikasi masalah, studi literatur, pengumpulan data, pelabelan sentimen, preprocessing data, pembobotan TF-IDF, reduksi dimensi menggunakan PCA, pembagian data (data splitting), proses klasifikasi menggunakan Logistic Regression, hingga evaluasi performa model. Kerangka kerja penelitian dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Kerja



2.2 Identifikasi Masalah

Ulasan pelanggan yang tersedia pada Google Reviews mengandung informasi penting mengenai pengalaman pelanggan terhadap produk dan layanan yang diberikan oleh perusahaan. Namun, jumlah ulasan yang terus bertambah menyebabkan proses analisis secara manual menjadi tidak efektif dan membutuhkan waktu yang lama. Selain itu, data ulasan berbentuk teks tidak terstruktur sehingga sulit digunakan secara langsung sebagai dasar pengambilan keputusan. Oleh karena itu, diperlukan suatu pendekatan analisis sentimen yang mampu mengklasifikasikan opini pelanggan ke dalam kategori sentimen positif dan negatif secara otomatis sehingga informasi yang terkandung dalam ulasan dapat dimanfaatkan secara lebih optimal sebagai bahan evaluasi dan peningkatan kualitas layanan.

2.3 Studi Literatur

Tahap studi literatur dilakukan untuk memperoleh landasan teoritis yang mendukung penelitian. Literatur yang digunakan meliputi konsep text mining, analisis sentimen, preprocessing teks, pembobotan TF-IDF, Principal Component Analysis (PCA), Logistic Regression, serta penelitian terdahulu yang relevan. Studi literatur bertujuan untuk memahami karakteristik metode yang digunakan, mengidentifikasi kelebihan dan keterbatasan penelitian sebelumnya, serta menentukan pendekatan yang sesuai untuk diterapkan dalam penelitian ini. Selain itu, studi literatur juga digunakan untuk menemukan research gap yang menjadi dasar pengembangan model analisis sentimen yang diusulkan.

2.4 Crawling Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa ulasan pelanggan yang diperoleh dari Google Reviews. Proses pengambilan data dilakukan menggunakan teknik web scraping dengan bantuan bahasa pemrograman Python. Data dikumpulkan dari sepuluh cabang d'Besto EBM yang berada di Kota Padang. Hasil proses scraping menghasilkan 679 ulasan pelanggan berbahasa Indonesia yang kemudian disimpan dalam format CSV untuk digunakan pada tahap pengolahan dan analisis selanjutnya.

2.5 Labelling Data

Tahap pelabelan data dilakukan untuk memberikan kategori sentimen pada setiap ulasan pelanggan. Pelabelan dilakukan secara manual berdasarkan makna keseluruhan isi ulasan dengan mengacu pada kecenderungan opini yang disampaikan pelanggan. Ulasan yang mengandung apresiasi, kepuasan, rekomendasi, atau pengalaman positif diberikan label positif. Sebaliknya, ulasan yang mengandung keluhan, kekecewaan, kritik, atau pengalaman negatif diberikan label negatif. Proses pelabelan dilakukan sebelum tahap pemodelan agar data memiliki kelas target yang akan digunakan dalam proses pembelajaran algoritma Logistic Regression. Hasil pelabelan kemudian digunakan sebagai variabel dependen dalam proses klasifikasi sentimen.

2.6 Pre-Processing

Tahap pre-processing merupakan tahapan yang penting dalam pengolahan teks yang bertujuan untuk mempersiapkan data ulasan pelanggan sebelum dilakukan proses analisis sentimen. Pada tahap ini data teks dibersihkan dan dinormalisasi agar dapat diproses oleh algoritma dengan lebih efektif. Adapun tahapan pada pre-processing yaitu cleaning, case folding, tokenizing, stopwords removal, normalisasi dan stemming.

2.7 Pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF)

Proses TF-IDF memiliki tujuan dalam mentransformasi data teks yang telah dipreproses menjadi representasi numerik dengan menghitung tingkat kepentingan setiap kata melalui dua komponen utama TF yang mengukur frekuensi kemunculan kata dalam suatu dokumen dan IDF yang memberikan bobot lebih tinggi pada kata-kata informatif yang jarang muncul di seluruh korpus. Dengan mengalikan kedua komponen tersebut dihasilkan nilai TF-IDF akhir yang mencerminkan relevansi serta pengaruh suatu kata terhadap isi dokumen secara keseluruhan sehingga kata-kata yang paling signifikan dapat diidentifikasi.

2.8 Metode Principal Component Analysis (PCA)

Setelah data teks ditransformasikan ke dalam bentuk numerik menggunakan TF-IDF, jumlah fitur yang dihasilkan menjadi sangat besar karena setiap kata direpresentasikan sebagai fitur tersendiri. Kondisi ini dapat meningkatkan kompleksitas komputasi dan memengaruhi performa model klasifikasi. Oleh karena itu, penelitian ini menerapkan Principal Component Analysis (PCA) untuk mereduksi dimensi fitur TF-IDF menjadi sejumlah komponen utama yang mampu mempertahankan sebagian besar informasi dari data asli. Setelah data teks ditransformasikan ke dalam bentuk numerik menggunakan TF-IDF, jumlah fitur yang dihasilkan menjadi sangat besar karena setiap kata direpresentasikan sebagai fitur tersendiri. Kondisi ini dapat meningkatkan kompleksitas komputasi dan memengaruhi performa model klasifikasi. Oleh karena itu, penelitian ini menerapkan Principal Component Analysis (PCA) untuk mereduksi dimensi fitur TF-IDF menjadi sejumlah komponen utama yang mampu mempertahankan sebagian besar informasi dari data asli. PCA bekerja dengan mentransformasikan fitur-fitur yang saling berkorelasi menjadi komponen baru yang tidak saling berkorelasi (principal components). Pemilihan jumlah komponen dilakukan berdasarkan nilai explained variance sehingga komponen yang dipilih mampu merepresentasikan sebagian besar variasi data. Pada penelitian ini diperoleh dua



komponen utama yang mampu menjelaskan 85,7% total variansi data sehingga digunakan sebagai representasi baru dalam proses klasifikasi sentimen menggunakan Logistic Regression.

2.9 Data Splitting

Splitting data bertujuan untuk membagi dataset menjadi dua bagian yaitu data latih dan data uji. Pembagian data dilakukan dengan rasio 80:20 adapun ketentuannya yaitu 80% data digunakan untuk melatih model sedangkan 20% sisanya digunakan untuk menguji kinerja model dalam memprediksi data baru. Proses ini memastikan bahwa model tidak hanya menghafal data yang sudah ada tetapi juga mampu melakukan generalisasi dengan baik saat menghadapi data yang belum pernah dilihat sebelumnya.

2.10 Logistic Regression

Tahap klasifikasi dilakukan menggunakan algoritma Logistic Regression. Algoritma ini digunakan untuk memodelkan hubungan antara fitur hasil reduksi PCA dengan kelas sentimen yang terdiri dari sentimen positif dan negatif. Logistic Regression menghasilkan nilai probabilitas yang kemudian dipetakan ke dalam kelas tertentu menggunakan fungsi sigmoid. Pada penelitian ini, fitur hasil reduksi PCA digunakan sebagai variabel independen, sedangkan label sentimen digunakan sebagai variabel dependen. Model dilatih menggunakan data latih dan dievaluasi menggunakan data uji untuk mengetahui kemampuan model dalam mengklasifikasikan ulasan pelanggan yang belum pernah dilihat sebelumnya. Evaluasi model dilakukan menggunakan confusion matrix serta metrik akurasi, precision, recall, dan F1-score.

3. HASIL DAN PEMBAHASAN

Tahap pembahasan dan hasil merupakan pengolahan seluruh data ulasan pelanggan d'Besto EBM yang telah dikumpulkan diolah menggunakan serangkaian metode Data Mining. Secara garis besar alur kerja dimulai dari tahap pre-processing untuk membersihkan data dari noise dilanjutkan dengan transformasi data menjadi bentuk numerik menggunakan teknik TF-IDF. Setelah data selesai dilakukan penerapan Principal Component Analysis (PCA) untuk mereduksi dimensi fitur guna meningkatkan efisiensi komputasi tanpa menghilangkan informasi krusial dalam data.

Data hasil reduksi dimensi tersebut diproses menggunakan algoritma Logistic Regression untuk mengklasifikasikan sentimen ke dalam kategori positif dan negatif. Hasil dari klasifikasi ini kemudian dievaluasi kinerjanya menggunakan confusion matrix untuk mengukur tingkat akurasi, presisi dan recall. Pada tahap akhir pola kepuasan pelanggan akan divisualisasikan untuk memberikan gambaran yang jelas bagi pihak manajemen mengenai aspek-aspek pelayanan dan kualitas produk yang paling berpengaruh terhadap persepsi konsumen.

3.1 Pre-Processing

Tahap preprocessing dilakukan untuk membersihkan dan menstandarisasi data ulasan pelanggan sebelum digunakan dalam proses klasifikasi sentimen. Proses ini bertujuan untuk mengurangi noise pada data teks sehingga informasi yang relevan dapat dipertahankan dan diproses secara optimal oleh algoritma klasifikasi. Tahapan preprocessing yang dilakukan meliputi cleaning, case folding, tokenizing, stopword removal, normalisasi, dan stemming. Hasil preprocessing menghasilkan representasi teks yang lebih terstruktur dan seragam dibandingkan data mentah. Tabel 1 menunjukkan contoh hasil preprocessing yang dilakukan terhadap data ulasan pelanggan.

Table 1. Data Pre-Processing

caption	label	text final
Lumayan, pengen banget makan ayam	Positif	cukup ingin banget makan ayam
Item Menu Baru: Minuman rasa markisa. Itu tidak penuh, dan ketika saya mencobanya, itu manis dan menyinggung tali vokal saya. Tidak direkomendasikan. CATATAN: Saya tidak kekurangan perubahan. Dia tidak mengatakan dia tidak memberi saya perubahan yang cukup, tetapi kasir itu gemuk.	Negatif	item menu minum markisa itu tidak penuh coba manis singgung tali vokal sayatidak direkomendasikan catatan kurang perubahandia ubah kasir gemuk
Rasa pedas di tingkat Sadas tidak konsisten, dan ada terlalu sedikit saus di Sadas.	Negatif	pedas tingkat sayap pedas tidak konsisten saus sayap pedas
Pelayanannya sangat buruk, tempatnya panas, kasirnya kasar dan mukanya jelek, tata kramanya juga buruk	Negatif	layan buruk tempat panas kasir kasar muka buruk tata krama buruk
Ayam yang sangat lezat, cepatlh, dbesto	Positif	ayam lezat cepat d besto
Cukup bagus	Negatif	bagus
Patty yang terbakar, ayam dingin, tidak ada kotak (biasanya ada satu) paha yang dipesan, yang tiba adalah payudara, toko tidak mengkonfirmasi terlebih dahulu. MENYELESAIKAN!	Negatif	patty gosong ayam dingin kotak paha pes payudara toko konfirmasi dahulumenyelesaikan
Cobalah saya membeli level 2 Sadas berharap untuk mendapatkan bagian sayap, tetapi saya diberi 2 potongan sayap kecil dan dua potong daging yang padat melihat layanannya di	Negatif	coba beli level sayap pedas harap sayap potong sayap potong



caption	label	text final
mana harganya memiliki kualitas yang sama dari makanan tidak lagi cocok		daging padat layan harga milik kualitas makan cocok
...		
Mohon maaf kalau memang tidak bisa menjaga kebersihan toilet, bilang saja tidak ada toilet.. Kalau begitu tolong bersihkan sampah di depan ya bu pemilik.. Wastafelnya juga kotor.. 😊	Negatif	mohon maaf jaga bersih toilet bilang toilet tolong sih sampah ya bu milik wastafel kotor menu relatif outlet
Menunya relatif hampir sama dengan outlet lain.		
Layanan ini responsif dan ramah, menu ini mudah dilihat, ada juga meja dan kursi yang dapat digunakan untuk makan di tempat, tetapi sayangnya tempat itu terlalu kecil dan panas jadi menurut saya cukup tidak nyaman berada di sana untuk waktu yang lama.	Negatif	layan responsif ramah menu mudah meja kursi makan sayang panas selesa
Bagus sekali	Positif	bagus

Berdasarkan Tabel 1, terlihat bahwa proses preprocessing berhasil menghilangkan karakter yang tidak diperlukan, menyeragamkan bentuk kata, serta mengubah kata berimbuhan menjadi kata dasar. Hasil akhir berupa kolom text_final digunakan sebagai input pada tahap pembobotan TF-IDF dan proses klasifikasi sentimen.

3.2 Transformasi TF-IDF

Data hasil preprocessing kemudian ditransformasikan ke dalam bentuk numerik menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). Metode ini digunakan untuk memberikan bobot pada setiap kata berdasarkan tingkat kemunculan kata dalam suatu dokumen dan keseluruhan korpus. Representasi TF-IDF menghasilkan matriks fitur berdimensi tinggi karena setiap kata unik direpresentasikan sebagai fitur tersendiri. Kondisi ini berpotensi meningkatkan kompleksitas komputasi dan memengaruhi performa model klasifikasi. Oleh karena itu, diperlukan teknik reduksi dimensi untuk menyederhanakan representasi data tanpa menghilangkan informasi penting.

3.3 Penerapan Principal Component Analysis

Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi data hasil pembobotan TF-IDF. Tujuan utama penerapan PCA adalah mengurangi jumlah fitur yang digunakan dalam proses klasifikasi sehingga model menjadi lebih efisien tanpa kehilangan sebagian besar informasi yang terkandung dalam data. Hasil PCA menunjukkan bahwa dua komponen utama pertama (PC1 dan PC2) mampu menjelaskan sebesar 85,7% total variansi data. Hal ini menunjukkan bahwa sebagian besar informasi yang terdapat pada fitur asli berhasil direpresentasikan oleh kedua komponen tersebut

Table 2. Penerapan PCA

Dokumen	PC1	PC2
D1	0,392	-0,061
D3	0,181	0,478
D8	-0,522	-0,034
D10	-0,441	-0,322

Tabel 2 menunjukkan hasil proyeksi beberapa dokumen ke dalam ruang fitur baru yang dibentuk oleh PC1 dan PC2. Nilai tersebut bukan menunjukkan sentimen positif atau negatif secara langsung, melainkan posisi masing-masing dokumen pada ruang fitur hasil reduksi dimensi. Dokumen yang memiliki posisi berdekatan menunjukkan kemiripan karakteristik fitur, sedangkan dokumen yang memiliki posisi berjauhan menunjukkan perbedaan karakteristik yang lebih besar. Dengan demikian, PCA berhasil menyederhanakan representasi data teks berdimensi tinggi menjadi dua komponen utama yang tetap mampu mempertahankan sebagian besar variasi informasi dari data asli.

3.4 Proses Metode Logistic Regression

Data hasil reduksi dimensi menggunakan PCA selanjutnya digunakan sebagai input pada algoritma Logistic Regression untuk melakukan klasifikasi sentimen. Logistic Regression bekerja dengan menghitung nilai probabilitas setiap dokumen terhadap kelas target melalui fungsi sigmoid. Nilai probabilitas yang dihasilkan berada pada rentang 0 hingga 1. Dalam penelitian ini, probabilitas yang lebih besar dari 0,5 dikategorikan sebagai sentimen positif, sedangkan probabilitas yang lebih kecil dari atau sama dengan 0,5 dikategorikan sebagai sentimen negatif.

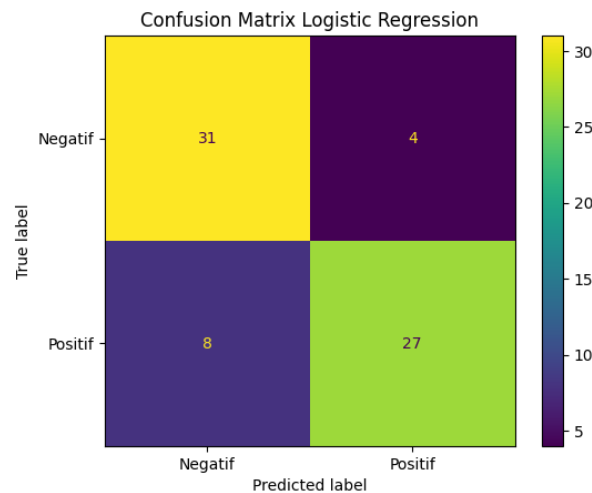
Tabel 3. Proses Metode Logistic Regression

Dokumen	z	Probabilitas	Kelas
D1	0,331	0.582	Positif
D3	0,659	0.659	Positif
D8	-0,556	0,364	Negatif
D10	-0,763	0,318	Negatif

Berdasarkan Tabel 3, terlihat bahwa nilai probabilitas berpengaruh langsung terhadap hasil klasifikasi sentimen. Dokumen dengan probabilitas di atas 0,5 diklasifikasikan sebagai sentimen positif, sedangkan dokumen dengan probabilitas di bawah 0,5 diklasifikasikan sebagai sentimen negatif. Hasil ini menunjukkan bahwa Logistic Regression mampu memanfaatkan fitur hasil reduksi PCA untuk membedakan ulasan pelanggan berdasarkan kecenderungan sentimennya.

3.5 Evaluasi Model

Evaluasi model dilakukan menggunakan confusion matrix untuk mengukur performa klasifikasi yang dihasilkan oleh Logistic Regression.



Gambar 2. Confusion Matrik

Berdasarkan hasil pengujian diperoleh sebanyak 70 data uji yang merupakan 20% dari total dataset. Hasil confusion matrix menunjukkan bahwa model berhasil mengklasifikasikan 58 data secara benar yang terdiri dari 31 True Negative (TN) dan 27 True Positive (TP). Sementara itu, terdapat 12 data yang mengalami kesalahan klasifikasi yang terdiri dari 4 False Positive (FP) dan 8 False Negative (FN). Berdasarkan nilai confusion matrix tersebut diperoleh metrik evaluasi sebagai berikut:

Accuracy = 82,86%

Precision = 87,10%

Recall = 77,14%

F1-Score = 81,82%

Nilai akurasi sebesar 82,86% menunjukkan bahwa model mampu melakukan klasifikasi sentimen dengan tingkat ketepatan yang baik. Nilai precision yang tinggi menunjukkan bahwa sebagian besar prediksi sentimen positif yang dihasilkan model sesuai dengan kondisi sebenarnya. Sementara itu, nilai recall menunjukkan kemampuan model dalam mengidentifikasi data yang benar-benar termasuk ke dalam kelas positif. Nilai F1-Score sebesar 81,82% menunjukkan keseimbangan yang baik antara precision dan recall. Hasil tersebut menunjukkan bahwa integrasi PCA dan Logistic Regression mampu menghasilkan model klasifikasi sentimen yang cukup efektif pada data ulasan pelanggan berbahasa Indonesia. Penerapan PCA berhasil mengurangi kompleksitas fitur TF-IDF dengan tetap mempertahankan sebagian besar informasi penting, sedangkan Logistic Regression mampu memanfaatkan representasi fitur tersebut untuk menghasilkan performa klasifikasi yang baik.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan metode Principal Component Analysis (PCA) dan Logistic Regression dalam melakukan klasifikasi sentimen terhadap ulasan pelanggan yang diperoleh dari Google Reviews. Berdasarkan hasil eksperimen, model yang dibangun menghasilkan nilai akurasi sebesar 82%, yang menunjukkan bahwa kombinasi PCA sebagai metode reduksi dimensi dan Logistic Regression sebagai metode klasifikasi mampu bekerja secara cukup efektif dalam menangani data teks yang bersifat tidak terstruktur. Penerapan PCA terbukti mampu menyederhanakan representasi fitur TF-IDF yang memiliki dimensi tinggi menjadi beberapa komponen utama dengan tingkat variansi yang tetap tinggi, yaitu sebesar 85,7%. Hal ini menunjukkan bahwa sebagian besar informasi penting dari data tetap dapat dipertahankan meskipun jumlah fitur telah direduksi secara signifikan. Reduksi dimensi ini juga membantu meningkatkan efisiensi proses komputasi pada tahap klasifikasi. Hasil evaluasi menggunakan confusion matrix menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mengklasifikasikan sentimen positif dan negatif, meskipun masih terdapat



sejumlah kesalahan klasifikasi yang ditunjukkan oleh nilai False Positive dan False Negative. Hal ini mengindikasikan bahwa masih terdapat ruang untuk peningkatan performa model, terutama dalam hal kualitas fitur dan metode pembelajaran yang digunakan. Penelitian ini memberikan gambaran bahwa integrasi PCA dan Logistic Regression dapat digunakan sebagai pendekatan awal yang efisien dalam analisis sentimen, namun masih memerlukan pengembangan lebih lanjut untuk mencapai performa yang lebih optimal.

REFERENCES

- [1] A. Pemasaran, "Affiliate Marketing and Smartphone Usage on Student Productivity," *Journal of Digital Marketing Research*, vol. 5, no. 2, pp. 1–9, 2025.
- [2] K. L. Keller, *Marketing Management: Strategic Brand Analysis*, Pearson Education, 2022.
- [3] B. Harto, A. Y. Rukmana, R. Tahir, and E. Waty, "Digital Business Transformation in the Era of Information Technology," *Journal of Business Innovation*, vol. 8, no. 1, pp. 45–56, 2023.
- [4] N. A. Rahma, A. Anggarani, and W. Prasetyoning, "The Role of Online Customer Review, Rating, and Brand Image on Purchase Decision," *Journal of Marketing Management*, vol. 12, no. 4, pp. 210–218, 2024, doi: 10.47191/jefms/v6-i6-17.
- [5] J. Nugraha, "Branding Strategy in the Digital Era," *Journal of Business Communication*, vol. 7, no. 1, pp. 33–41, 2024, doi: 10.4018/978-1-6684-8351-0.ch008
- [6] A. S. Sitio and F. A. Sianturi, "Machine Learning Application in Internet Usage Behavior Analysis," *Journal of Information Systems*, vol. 2, no. 1, pp. 46–51, 2024.
- [7] APJII, "Indonesia Internet Usage Report 2023," Association of Indonesian Internet Service Providers, 2023. <https://apjii.or.id>
- [8] S. Mishra et al., "Principal Component Analysis," *International Journal of Learning and Research*, vol. 10, no. 1, pp. 1–10, 2022. <https://doi.org/10.5455/ijlr.20170415115235>
- [9] D. N. Ardelia, H. D. Arifin, S. Daniswara, and A. P. Sari, "Mobile Phone Price Classification Using Logistic Regression Algorithm," *Journal of Computer Science and Informatics*, vol. 4, no. 1, pp. 37–43, 2024, doi: 10.1109/ISMSIT52890.2021.9604550.
- [10] I. Rahmawati and T. R. Fitriani, "Sentiment Analysis Using Logistic Regression on Lion Air Airline Reviews," *Journal of Information Technology*, vol. 1, no. 1, pp. 11–16, 2023, doi: 10.58776/jriti.v1i1.60
- [11] S. H. Pratama, "Big Data Analytics and Predictive Modeling in Industry 4.0," *Journal of Data Science and Analytics*, vol. 6, no. 2, pp. 55–63, 2023.
- [12] P. Yayasan and K. Menulis, "Application of Artificial Intelligence in Data Science," *Journal of AI Research*, vol. 3, no. 1, pp. 22–30, 2025.
- [13] S. D. Anggita and F. F. Abdulloh, "Sentiment Analysis Optimization Using PSO and Information Gain Feature Selection with SVM," *Journal of Machine Learning Applications*, vol. 4, no. 1, pp. 52–57, 2023, doi: 10.52158/jacost.v4i1.524
- [14] Z. Shen, "Feature Selection Using Fuzzy Methods and Meta Learning," *Journal of Computational Intelligence*, vol. 9, no. 2, pp. 88–97, 2022, doi: 10.1109/FUZZ48607.2020.9177769.
- [15] A. K. Jean, M. Diarra, B. A. Bakary, G. Pierre, and A. K. Jérôme, "Hybrid CNN-SVM and PCA-SVM Approaches for Cocoa Beans Classification," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 100–110, 2022. <https://doi.org/10.14569/IJACSA.2022.0130927>
- [16] H. Rachmi, "PCA and Genetic Algorithm for Sentiment Analysis Using SVM," *Journal of Data Mining and Knowledge Discovery*, vol. 5, no. 2, pp. 97–104, 2022.
- [17] K. Srijiaranon, "Hybrid PCA-EMD-LSTM Framework for Stock Market Prediction with Sentiment Analysis," *Applied Sciences*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.3390/app122110823.
- [18] Z. Song, X. Gong, C. Zhang, and C. Yu, "Investor Sentiment Based on Scaled PCA Method for Stock Market Volatility Prediction," *International Review of Economics & Finance*, vol. 83, pp. 528–545, 2023. <https://doi.org/10.1016/j.iref.2022.10.007>
- [19] E. O. Omuya, G. Okeyo, and M. Kimwele, "Sentiment Analysis on Social Media Tweets Using Dimensionality Reduction and NLP Techniques," *Engineering Reports*, vol. 14, no. 2, pp. 1–14, 2022. <https://doi.org/10.1002/eng2.12579>
- [20] S. Srivastava, C. Chakraborty, and M. Kanti, "Leveraging Machine Learning and Dimensionality Reduction for Sentiment Analysis," *Measurement: Sensors*, vol. 33, pp. 101182, 2024. <https://doi.org/10.1016/j.measen.2024.101182>
- [21] Y. Zhang and L. Wang, "Text Classification Using Logistic Regression and TF-IDF Feature Selection," *Journal of Big Data Research*, vol. 11, no. 3, pp. 101–112, 2024, doi: 10.47467/visa.v4i3.2835.
- [22] R. Kumar, "Dimensionality Reduction Techniques in Text Mining: A Review," *Journal of Artificial Intelligence Review*, vol. 15, no. 2, pp. 55–70, 2023, doi: 10.4018/978-1-5225-0489-4.ch003.