Analisis Ulasan Konsumen sebagai Data Non-Keuangan dalam Sistem Informasi Akuntansi

Lady Agustin Fitriana[1*], Muhammad Fahmi Julianto[2], Rizka Dahlia [3], Muhammad Rifqi Firdaus [4], Agung Fazriansyah[5]
[1,2,3,4,5] Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Senen, Jakarta Pusat, Indonesia
e-mail: lady.lag@bsi.ac.id, fahmi.fjl@bsi.ac.id, rizka.rzl@bsi.ac.id, muhammad.mku@bsi.ac.id, agung.fzr@bsi.ac.id

Artikel Info : Diterima : 10-04-2025 | Direvisi : 10-05-2025 | Disetujui : 10-06-2025

Abstrak - Di era digital saat ini, ulasan pengguna pada platform e-commerce seperti Shopee telah menjadi salah satu sumber informasi non-keuangan yang penting dalam menilai persepsi konsumen terhadap produk maupun layanan. Informasi ini memiliki nilai strategis dalam sistem informasi akuntansi, khususnya dalam mendukung pengambilan keputusan berbasis data pelanggan. Namun, tantangan utama yang dihadapi adalah besarnya volume data ulasan yang tidak terstruktur. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi metode yang efektif dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Shopee sebagai data non-keuangan yang dapat digunakan dalam sistem informasi akuntansi manajerial. Penelitian ini memanfaatkan kombinasi teknik *Natural Language Processing (NLP)* dan algoritma *K-Nearest Neighbors (KNN)*. Data ulasan dikumpulkan melalui proses crawling dari Google Play Store menggunakan pustaka google-play-scraper. Data tersebut kemudian diproses melalui serangkaian tahapan NLP seperti case folding, tokenization, normalisasi, penghapusan stopword, dan stemming. Untuk ekstraksi fitur, digunakan metode TF-IDF dan Cosine Similarity untuk menghasilkan representasi vektor yang sesuai dengan kebutuhan klasifikasi. Model KNN digunakan untuk mengklasifikasikan sentimen berdasarkan data latih, dengan pengujian pada berbagai nilai n_neighbors. Hasil penelitian menunjukkan bahwa model dengan n_neighbors = 9 menghasilkan akurasi 88%, presisi 85%, recall 86%, dan *F1-score* 85%.Temuan ini menunjukkan bahwa kombinasi NLP dan KNN efektif dalam mengklasifikasikan sentimen ulasan pengguna, serta berpotensi besar untuk diterapkan sebagai bagian dari sistem informasi akuntansi guna memperkuat analisis non-keuangan dalam mendukung evaluasi kinerja penjualan dan pengambilan keputusan manajerial.

Kata Kunci : Sentimen Analisis, Natural Language Processing (NLP), K-Nearest Neighbors (KNN)

Abstrac - *In today's digital era, user reviews on e-commerce platforms such as Shopee have become a valuable source of non-financial information for assessing consumer perceptions of products and services. This information holds strategic significance in accounting information systems, particularly in supporting data-driven managerial decision-making. However, a major challenge lies in processing the large volume of unstructured review data. Therefore, this study aims to explore an effective method for classifying the sentiment of user reviews for the Shopee application as non-financial data that can be utilized within managerial accounting information systems.This study employs a combination of Natural Language Processing (NLP) techniques and the K-Nearest Neighbors (KNN) algorithm. Review data was collected through web crawling from the Google Play Store using the google-play-scraper library. The data was then preprocessed using a series of NLP steps, including case folding, tokenization, normalization, stopword removal, and stemming. For feature extraction, TF-IDF and Cosine Similarity methods were used to generate suitable vector representations for classification. The KNN model was applied to classify sentiment based on training data, with testing conducted across various n_neighbors values. The results show that the model with n_neighbors = 9 achieved an accuracy of 88%, a precision of 85%, a recall of 86%, and an F1-score of 85%.These findings demonstrate that the integration of NLP and KNN is effective in classifying user sentiment and has strong potential for application within accounting information systems. By incorporating such non-financial data, businesses can enhance performance evaluation and support managerial decision-making processes based on consumer feedback analytics.*

Keywords : Analysist Sentiment, Natural Language Processing (NLP), K-Nearest Neighbors (KNN)

PENDAHULUAN

Kemajuan teknologi digital diyakini telah memberikan efek yang substansial pada berbagai area kehidupan, termasuk dalam ranah perdagangan. Kemunculan e-commerce diungkapkan sebagai salah satu manifestasi transformasi digital yang paling menonjol dalam beberapa tahun terakhir (Muqoddas et al., 2020)(Creazza et al., 2023). Shopee, sebagai salah satu platform e-commerce terkemuka di Asia Tenggara, khususnya di Indonesia, menyediakan ruang bagi konsumen untuk memberikan ulasan terhadap produk yang mereka beli (Sihombing et al., 2021). Keberadaan ulasan tersebut dilaporkan penting sebagai sumber informasi

yang menggambarkan persepsi dan kepuasan pelanggan terhadap suatu produk atau layanan. (Masripah & Utami, 2020)(Dwiki et al., 2021).

Namun demikian, ulasan konsumen yang tersedia dalam jumlah besar dan bersifat tidak terstruktur menimbulkan tantangan tersendiri dalam hal pengolahan dan pemaknaan data. Dalam konteks ini, analisis sentimen merupakan cabang dari penambangan teks (text mining) yang bertujuan untuk mengidentifikasi dan mengklasifikasikan opini pengguna terhadap suatu entitas ke dalam kategori sentimen positif, negatif, atau netral(Agustine Fitrana et al., 2024). Penerapan teknik ini dilaporkan seringkali terjadi dalam berbagai bidang, termasuk evaluasi produk, penilaian layanan, dan analisis opini publik (Wankhade et al., 2022). Salah satu kesulitan utama dalam analisis sentimen adalah bagaimana mengelola data teks yang tidak terstruktur serta memperoleh informasi yang relevan dari bahasa alami yang digunakan oleh pengguna (Kusumaningrum et al., 2023).

Untuk mengatasi permasalahan tersebut, Natural Language Processing (NLP) atau pemrosesan bahasa alami merupakan sebuah bidang yang menggabungkan linguistik, ilmu komputer, dan kecerdasan buatan, yang memungkinkan mesin untuk memahami, menginterpretasi, dan menghasilkan bahasa manusia(Patil et al., 2023)(Singh & Mahmood, 2021)(Dahiya et al., 2023). Tahapan dalam NLP meliputi proses pembersihan data, tokenisasi, stemming, dan penghapusan kata-kata umum (stopword) untuk mempersiapkan data agar dapat dianalisis secara sistematis (Khurana et al., 2023).

Dalam analisis sentimen berbasis NLP, diperlukan algoritma klasifikasi untuk mengelompokkan data ulasan ke dalam kategori sentiment (Yang et al., 2020). Salah satu algoritma yang banyak digunakan adalah K-Nearest Neighbors (KNN), yaitu metode pembelajaran terawasi (supervised learning) yang melakukan klasifikasi berdasarkan kemiripan terhadap sejumlah tetangga terdekat (Ernawan et al., 2022). Algoritma KNN dikenal karena kesederhanaannya serta efektivitasnya dalam menangani masalah klasifikasi berbasis teks (Syarifuddin, 2020).

Beberapa penelitian sebelumnya telah menunjukkan bahwa kombinasi antara NLP dan algoritma KNN mampu menghasilkan kinerja yang baik dalam tugas klasifikasi sentimen. Syafrizal dkk menganalisis ulasan pengguna PLN Mobile menggunakan metode text mining dengan algoritma Naive Bayes dan K-Nearest Neighbors (KNN), serta validasi menggunakan K-Fold Cross Validation. Hasil pelabelan menunjukkan bahwa 69,97% ulasan memiliki sentimen positif, 12,27% netral, dan 17,77% negatif. Algoritma Naive Bayes memberikan akurasi tertinggi, yaitu sebesar 77,69%. (Syafrizal et al., 2023). Penelitian selanjutnya dikemukakan oleh Deta Kiran dan Al Faraby yang menerapkan algoritma K-Nearest Neighbor (KNN) dalam menganalisis ulasan produk kecantikan. Hasil penelitian mengindikasikan bahwa tingkat akurasi tertinggi yang diperoleh mencapai 77,69%, yang dicapai dengan menggunakan algoritma Naive Bayes. 71% dengan nilai k=50 dan menggunakan 76 fitur hasil seleksi (Deta Kirana & Al Faraby, 2021). Sementara itu, Luban Abdi Susanto membandingkan performa algoritma K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM) terhadap ulasan aplikasi Polri Super App. Studi ini menunjukkan bahwa SVM dengan kernel linear menghasilkan akurasi tertinggi sebesar 89,67%, diikuti oleh SVM dengan kernel RBF dan KNN. Terakhir, penelitian oleh Elik menggunakan algoritma KNN dan metode TF-IDF dalam analisis ulasan produk hijab instan di marketplace, dengan mengadopsi pendekatan Natural Language Processing (NLP). Hasil penelitian mengungkapkan bahwa pendekatan berbasis NLP menghasilkan akurasi sebesar 76,92%, presisi 80,00%, dan recall 74,07%, yang terbukti lebih unggul dibandingkan dengan pendekatan tanpa menggunakan NLP, yang hanya mencapai akurasi 69,23% (Muktafin et al., 2020).

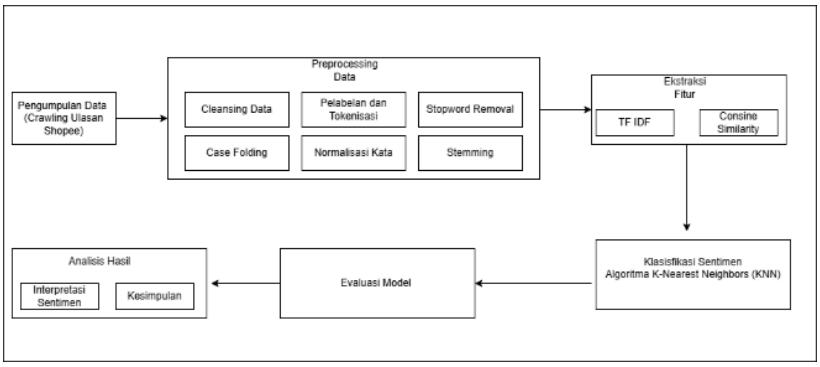
Berdasarkan temuan dari berbagai penelitian tersebut, penelitian ini bertujuan membangun model klasifikasi sentimen berbasis NLP dan KNN untuk mengelompokkan ulasan pengguna aplikasi Shopee ke dalam kategori sentimen positif, negatif, dan netral. Penelitian ini mencakup beberapa tahapan: pengumpulan data dari Google Play Store, praproses data menggunakan teknik NLP, klasifikasi menggunakan algoritma KNN, dan evaluasi performa model menggunakan metrik akurasi, presisi, dan recall. Dengan mengembangkan hipotesis bahwa kombinasi NLP dan KNN dapat meningkatkan akurasi dalam klasifikasi sentimen, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem analisis sentimen pada sektor e-commerce, khususnya dalam meningkatkan pemahaman terhadap opini konsumen di platform Shopee..

METODE PENELITIAN

Penelitian ini menganalisis sentimen ulasan Shopee di Play Store menggunakan Natural Language Processing (NLP) dan K-Nearest Neighbors (KNN). Tahapan penelitian dimulai dengan pengumpulan data melalui proses crawling, diikuti dengan tahap preprocessing yang mencakup pembersihan teks, normalisasi, penghapusan kata berhenti (stop-word removal), tokenisasi, dan stemming untuk menyiapkan data sebelum dilakukan analisis lebih lanjut.Setelah itu, dilakukan transformasi data, termasuk penerjemahan ke bahasa Inggris dan labeling sentimen sebagai positif atau negatif. Selanjutnya, dalam ekstraksi fitur, digunakan TF-IDF untuk mengubah teks menjadi angka dan Cosine Similarity untuk mengukur kedekatan antar teks. Proses klasifikasi sentimen dilakukan dengan

KNN, yang menentukan apakah suatu ulasan positif atau negatif berdasarkan data latih. Model kemudian dievaluasi dengan akurasi, recall, dan precision, serta divisualisasikan menggunakan menggunakan wordcloud. Hasil analisis menentukan dominasi sentimen positif atau negatif dalam ulasan Shopee serta mengevaluasi efektivitas NLP dan KNN dalam klasifikasi sentimen.

Adapun tahapan penelitian dapat dideskripsikan seperti Gambar 1.



Gambar 1 Tahapan Penelitian

Tahapan penelitian membantu peneliti untuk mengatur cara mengumpulkan data, menganalisisnya, dan menyusun kesimpulan berdasarkan informasi yang ditemukan.

1. Pengumpulan Data
   Proses pengumpulan data melalui crawling dengan menggunakan pustaka google-play-scraper untuk mengekstrak ulasan pengguna Shopee dari Google Play Store secara otomatis. Proses ini mengambil data berdasarkan ID aplikasi (com.shopee.id), dengan filter bahasa Indonesia dan wilayah Indonesia, serta mengurutkan ulasan dari yang terbaru. Hasilnya ditemukan 1.000 ulasan yang dikonversi ke DataFrame Pandas dan disimpan dalam format CSV, kemudian langsung ditampilkan agar dapat dianalisis dengan mudah melalui Google Colab seperti yang ditampilkan pada Gambar 2.

```
from google_play_scraper import app
import pandas as pd
import numpy as np
from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'com.shopee.id',
    lang='id',
    country='id',
    sort=Sort.MOST_RELEVANT,
    count=1000,
    filter_score_with=None
)
df_shopee = pd.DataFrame(np.array(result), columns=['review'])
df_shopee = df_shopee.join(pd.DataFrame(df_shopee.pop('review').tolist()))
df_shopee_cod = df_shopee[df_shopee['content'].str.contains('COD', case=False)]
df_shopee_cod.head()
```

Gambar 2 Proses *Crawling*

2. Preprocessing Data
   Dalam penelitian ini, Natural Language Processing (NLP) digunakan untuk membersihkan dan menyiapkan data teks sebelum dianalisis menggunakan algoritma K-Nearest Neighbors (KNN). Proses preprocessing data terdiri dari beberapa tahapan utama:
   1) *Case Folding*
      Tahapan case folding adalah proses mengubah semua huruf dalam tks menjadi huruf kecil (lowercase) atau huruf besar (uppercase) (Kosasih & Alberto, 2021). Dalam hal ini semua teks yang ditemukan dirubah menjadi huruf kecil untuk menyeragamkan teks dan menghindari perbedaan interpretasi karena perbedaan huruf kapitalisasi
   2) *Cleansing*
      Cleansing data, yang juga dikenal sebagai pembersihan data, adalah proses untuk mengidentifikasi serta memperbaiki (atau menghapus) data yang salah, tidak akurat, tidak lengkap, tidak relevan, atau rusak dalam sebuah dataset, guna memastikan kualitas data yang lebih baik untuk analisis atau pemrosesan lebih lanjut (Ferlay et al., 2021). Dataset yang telah diproses kemudian dilakukan Proses ini melibatkan penghapusan karakter-karakter yang tidak

relevan seperti tanda baca, angka, dan karakter tunggal, serta menghilangkan whitespace yang berlebihan, untuk memastikan teks yang bersih dan terstruktur dengan baik sebelum dilakukan analisis lebih lanjut.

3) Pelabelan Data dan Tokenisasi
Pelabelan data dalam analisis sentimen adalah proses memberikan label pada data ulasan atau teks untuk mengklasifikasikannya ke dalam kategori sentimen tertentu, seperti positif, negatif, atau netral, guna mempermudah pemahaman dan analisis terhadap persepsi pengguna atau opini yang terkandung dalam teks tersebut (Syafrizal et al., 2023).. Tokenisasi adalah proses dekomposisi teks menjadi unit-unit linguistik yang lebih kecil, yang disebut token (Asaad & Abdulhakim, 2021). Dalam penelitian ini, token merepresentasikan kata-kata individual dalam ulasan pengguna. Proses tokenisasi dilakukan dengan memanfaatkan fungsi word_tokenize yang disediakan oleh library Natural Language Toolkit (NLTK) dalam Python. NLTK adalah toolkit komprehensif untuk pemrosesan bahasa alami. Untuk mengintegrasikan fungsi ini ke dalam alur kerja preprocessing, sebuah fungsi pembungkus (word_tokenize_wrapper (text)) didefinisikan. Fungsi ini menerima teks sebagai input dan mengembalikan daftar token yang dihasilkan.
4) Normalisasi Kata
Normalisasi kata bertujuan untuk menyeragamkan representasi teks dengan mengganti kata-kata non-standar atau varian kata (misalnya, kata slang, singkatan) dengan bentuk kata yang baku. Langkah ini penting untuk mengurangi variasi leksikal dan meningkatkan konsistensi data.
5) Stopword Removal
Penghapusan stopword adalah teknik untuk menghapus kata-kata umum seperti "dan", "atau", "adalah", yang dianggap memiliki sedikit atau bahkan tidak ada kontribusi signifikan terhadap analisis sentimen, sehingga fokus analisis dapat lebih diarahkan pada kata-kata yang lebih penting dan relevan. unit-unit linguistik yang lebih kecil, yang disebut token (Asaad & Abdulhakim, 2021). Dalam penelitian ini, stopword dari Bahasa Indonesia dan Inggris dihilangkan menggunakan stopword list yang disediakan oleh NLTK, yang dikombinasikan dengan stopword list tambahan yang dikelola secara manual dan disimpan dalam file stopwords.csv. Kombinasi ini bertujuan untuk mencakup stopword yang relevan dengan konteks spesifik dari data ulasan pengguna. Fungsi stopword_removal(Review) mengimplementasikan proses ini. Fungsi ini memfilter daftar token, menghilangkan token apa pun yang ada dalam stopword list.
6) *Stemming*
Stemming adalah proses untuk mereduksi kata-kata ke bentuk dasarnya (kata dasar atau root) (Agustine Fitrana et al., 2024). Dalam penelitian ini, stemming dilakukan menggunakan library Sastrawi untuk Bahasa Indonesia. Fungsi stemming (Review) menerapkan stemming pada setiap token dan mengembalikan hasil berupa gabungan kata dasar.

3. Ekstraksi Fitur
1) TF-IDF
Tahapan TF-IDF *(Term Frequency-Inverse Document Frequency)* dilakuakn untuk mengubah teks menjadi representasi numerik yang dapat dipahami oleh model machine learning kemudian menghitung bobot TF-IDF untuk setiap kata dalam ulasan. TF-IDF memberikan bobot yang tinggi pada kata-kata yang sering muncul dalam dokumen tertentu tetapi jarang muncul di seluruh korpus, sehingga membantu menyoroti kata-kata yang penting untuk membedakan ulasan.
2) Consine Similarity
Setelah teks diubah menjadi representasi TF-IDF, kode menghitung cosine similarity antara setiap pasangan ulasan. Cosine similarity mengukur kesamaan antara dua vektor dalam ruang multidimensi. Dalam konteks ini, ini mengukur seberapa mirip ulasan satu sama lain berdasarkan bobot TF-IDF dari kata-kata. Matriks cosine similarity ini kemudian digunakan sebagai fitur untuk model KNN.

4. Algoritma K-Nearest Neighbors (KNN)
Algoritma KNN diterapkan pada data ulasan Shopee yang telah dibersihkan dan dihitung kemiripannya menggunakan Cosine Similarity. Data tersebut dibagi menjadi dua set, yaitu set pelatihan dan set

pengujian, dengan pembagian menggunakan metode train_test_split dengan rasio 80:20, 70:30, dan 60:40. Selanjutnya, model KNN dilatih menggunakan KNeighbors Classifier dari Scikit-learn untuk memprediksi sentimen (positif atau negatif) pada data pengujian. Setelah prediksi dilakukan, model dievaluasi menggunakan berbagai metrik seperti akurasi, precision, recall, f1-score, dan confusion matrix untuk menilai kinerja model dalam mengklasifikasikan sentimen.

5. Evaluasi Model
   Evaluasi model KNN dilakukan dengan menggunakan beberapa metrik, yaitu Confusion Matrix untuk menunjukkan jumlah prediksi yang benar dan salah pada masing-masing kelas (positif dan negatif), Classification Report untuk memberikan detail tentang precision, recall, dan f1-score setiap kelas, serta Cross-Validation menggunakan cross_val_score untuk melakukan validasi silang sebanyak 10 kali guna mengukur performa model pada berbagai subset data.

HASIL PENELITIAN DAN PEMBAHASAN

1. Pengumpulan Data

Hasil dari crawling mencakup informasi seperti nama pemberi ulasan, nilai bintang yang diberikan, waktu ulasan dikirim, dan isi review ulasan. Data ini diambil berdasarkan ID aplikasi com.shopee.id, dengan filter bahasa Indonesia dan wilayah Indonesia, serta pengurutan ulasan terbaru. Setelah itu, data dikonversi menjadi DataFrame Pandas untuk memudahkan analisis dan disimpan dalam format CSV. Data yang telah disimpan kemudian ditampilkan di Google Colab agar dapat diverifikasi dan dianalisis lebih lanjut..

Berikut merupakan tampilan hasil crawling menggunakan google-play-scraper seperti yang ditampilkan pada Gambar 3.

	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion
8	e06142ff-ac08-4fc9-96db-9017c0fa998b	Cucun	https://play-lh.googleusercontent.com/a-/ALV-U...	Aplikasi sopee ini sangat bagus. harga terjaka...	5	689	3.47.40	2025-04-21 09:35:06	Halo Kak Cucun, terima kasih ya udah jadi pela...	2025-04-17 06:11:40	3.47.40
25	3346311d-011b-4c9f-890c-0fde5d407733	Florentina Della	https://play-lh.googleusercontent.com/a-/ALV-U...	fitur COD tidak bisa lagi dipakai. Mana jasa k...	1	7	3.48.31	2025-04-29 03:38:38	Hai kak Florentina Della. Maaf ya buat kamu ng...	2025-04-29 04:35:31	3.48.31
32	fa484fa0-8dad-4e65-8c73-01934b04d215	Aprilia Diany	https://play-lh.googleusercontent.com/a/ACg8oc...	Saya udh bertahun tahun pake shopee tapi kecew...	3	0	3.49.35	2025-05-01 14:23:32	Hi ka Aprilia Diany, aku bantu infoin ya kak k...	2025-05-01 15:28:28	3.49.35
43	3cd7e2e1-a460-4e97-a0ee-a05171e0f481	Tegoeh Widodo	https://play-lh.googleusercontent.com/a/ACg8oc...	Alhamdulillah selama saya belanja di shopee ba...	5	31	3.48.28	2025-04-19 08:07:49	Hai kak. Makasih buat penilaian dan feedback y...	2022-11-03 01:23:45	3.48.28
48	cae9656d-8bb4-4591-8f1d-a12c987c1a51	EVLT 12	https://play-lh.googleusercontent.com/a-/ALV-U...	bintang satu untuk keamanan sistem. udah 4 kal...	1	3	3.47.40	2025-04-29 07:53:27	Hi kak EVLT 12, maaf ya udh buat resah terkait...	2025-04-29 08:38:45	3.47.40

Gambar 3 Hasil *Crawling* Ulasan Pengguna

2. Preprocessing Data

Pada tahapan preprocessing, dilakukan beberapa langkah untuk mempersiapkan data ulasan, yaitu case folding untuk mengubah semua teks menjadi huruf kecil, pelabelan dataset untuk memberi label sentimen pada ulasan, tokenisasi untuk memecah teks menjadi kata-kata atau token, normalisasi untuk mengganti variasi kata dengan bentuk standar, stopword removal untuk menghapus kata-kata yang tidak relevan, dan stemming untuk mengubah kata menjadi bentuk dasar. Proses ini bertujuan untuk membersihkan dan menyusun data agar siap untuk analisis lebih lanjut, seperti analisis sentimen atau klasifikasi teks.

a) Case Folding
   Proses ini membersihkan teks ulasan pengguna dengan beberapa tahapan preprocessing, termasuk menghapus mention, hashtag, URL, angka, tanda baca, dan emoji. Proses juga mencakup penghilangan karakter yang muncul lebih dari dua kali berturut-turut dan mengubah teks menjadi huruf kecil untuk konsistensi. Setelah itu, data yang telah dibersihkan diurutkan berdasarkan waktu ulasan dan disimpan dalam format CSV untuk dianalisis lebih lanjut.

Tabel 1. Proses *Case Folding*

Ulasan	3/lebih	Case Folding
Onclick BCA tiba-tiba hilang, pas mau ditambahkan lagi keterangannya "Nomor ini sudah terdaftar". Dicek di BCA memang masih tertaut sama Shopee dan ShopeePay, tapi di aplikasi Shopee-nya nggk ada 😭😔😤 Mana lagi urgent!..	Onclick BCA tiba-tiba hilang, pas mau ditambahkan lagi keterangannya "Nomor ini sudah terdaftar". Dicek di BCA memang masih tertaut sama Shopee dan ShopeePay, tapi di aplikasi Shopee-nya nggk ada Mana lagi urgent!	onclick bca tiba-tiba hilang, pas mau ditambahkan lagi keterangannya 'nomor ini sudah terdaftar'. dicek di bca memang masih tertaut sama shopee dan shopeepay, tapi di aplikasi shopee-nya nggk ada mana lagi urgent!.

b) Cleansing

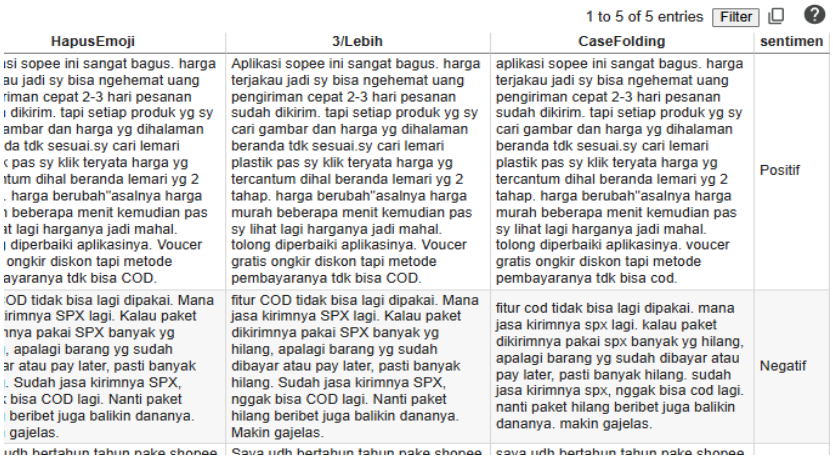
Tahapan ini berfungsi untuk membersihkan teks ulasan dengan menghapus karakter-karakter yang tidak relevan. Fungsi cleansing menghilangkan spasi ekstra, tanda baca (seperti ?, !, $, dll.), angka, serta kata yang hanya terdiri dari satu huruf. Selain itu, fungsi ini juga menyederhanakan spasi menjadi satu spasi tunggal.

Tabel 2. Proses *Cleansing Data*

Ulasan	*Cleansing Data*
aplikasi baguss walau kadang lemott . tolong dong biaya penangan jangan naik teruss yg awalnya tf bank cuma biaya 1.000 sekarang masak 2.500 makin mahaallllll plis dong pee shopeeee	aplikasi baguss walau kadang lemott tolong dong biaya penangan jangan naik teruss yg awalnya tf bank cuma biaya sekarang masak makin mahaallllll plis dong pee shopeeee

c) Pelabelan Dataset dan Tokenisasi

Pelabelan dataset dilakukan secara manual dengan mengkategorikan ulasan ke dalam kelas sentimen positif, negatif, atau netral. Proses pelabelan ini dilakukan dengan bantuan ahli bahasa Indonesia dan melibatkan setidaknya dua orang untuk memastikan akurasi pelabelan. Hasil pelabelan dapat ditunjukkan pada Gambar 4.



1 to 5 of 5 entries Filter

HapusEmoji	3/Lebih	CaseFolding	sentimen
ısi sopee ini sangat bagus. harga au jadi sy bisa ngehemat uang rimanan cepat 2-3 hari pesanan dikirim. tapi setiap produk yg sy ambar dan harga yg dihalaman da tdk sesuai.sy cari lemari pas sy klik teryata harga yg itum dihal beranda lemari yg 2 . harga berubah"asalnya harga beberapa menit kemudian pas t lagi harganya jadi mahal. diperbaiki aplikasinya. Voucer ongkir diskon tapi metode ayaranya tdk bisa COD.	Aplikasi sopee ini sangat bagus. harga terjakau jadi sy bisa ngehemat uang pengiriman cepat 2-3 hari pesanan sudah dikirim. tapi setiap produk yg sy cari gambar dan harga yg dihalaman beranda tdk sesuai.sy cari lemari plastik pas sy klik teryata harga yg tercantum dihal beranda lemari yg 2 tahap. harga berubah"asalnya harga murah beberapa menit kemudian pas sy lihat lagi harganya jadi mahal. tolong diperbaiki aplikasinya. Voucer gratis ongkir diskon tapi metode pembayaranya tdk bisa COD.	aplikasi sopee ini sangat bagus. harga terjakau jadi sy bisa ngehemat uang pengiriman cepat 2-3 hari pesanan sudah dikirim. tapi setiap produk yg sy cari gambar dan harga yg dihalaman beranda tdk sesuai.sy cari lemari plastik pas sy klik teryata harga yg tercantum dihal beranda lemari yg 2 tahap. harga berubah"asalnya harga murah beberapa menit kemudian pas sy lihat lagi harganya jadi mahal. tolong diperbaiki aplikasinya. voucer gratis ongkir diskon tapi metode pembayaranya tdk bisa cod.	Positif
OD tidak bisa lagi dipakai. Mana irimnya SPX lagi. Kalau paket nya pakai SPX banyak yg , apalagi barang yg sudah ar atau pay later, pasti banyak . Sudah jasa kirimnya SPX, bisa COD lagi. Nanti paket beribet juga balikin dananya. gajelas.	fitur COD tidak bisa lagi dipakai. Mana jasa kirimnya SPX lagi. Kalau paket dikirimnya pakai SPX banyak yg hilang, apalagi barang yg sudah dibayar atau pay later, pasti banyak hilang. Sudah jasa kirimnya SPX, nggak bisa COD lagi. Nanti paket hilang beribet juga balikin dananya. Makin gajelas.	fitur cod tidak bisa lagi dipakai. mana jasa kirimnya spx lagi. kalau paket dikirimnya pakai spx banyak yg hilang, apalagi barang yg sudah dibayar atau pay later, pasti banyak hilang. sudah jasa kirimnya spx, nggak bisa cod lagi. nanti paket hilang beribet juga balikin dananya. makin gajelas.	Negatif
udh bertahun tahun pake shopee	Saya udh bertahun tahun pake shopee	saya udh bertahun tahun pake shopee	

Sumber: Peneliti 2025

Gambar 4. Hasil Pelabelan

Setelah tahap case folding, di mana semua teks diubah menjadi huruf kecil, kalimat-kalimat akan diproses lebih lanjut dengan menguraikannya menjadi token-token atau kata-kata. Proses tokenisasi ini akan dibantu menggunakan library NLTK untuk memudahkan pemisahan teks menjadi unit-unit yang lebih kecil, sehingga data dapat diproses untuk analisis selanjutnya

CaseFolding	sentimen	Tokenizing
aplikasi semakin lemot, jaringan bagus dengan ...	Negatif	[aplikasi, semakin, lemot, ,, jaringan, bagus,...
sangat suka . aplikasi belanja terbaik . pilih...	Positif	[sangat, suka, ., aplikasi, belanja, terbaik, ...

Sumber: Peneliti 2025

Gambar 5. Hasil Tokenisasi

Dari Gambar 5 menunjukkan bahwa hasil tokenisasi mengubah data yang sebelumnya telah dibersihkan menjadi kata-kata terpisah yang nantinya memudahkan ekstraksi fitur dan evaluasi model

d) Normalisasi Kata

Tahap normalisasi kata bertujuan untuk mengubah penggunaan kata tidak baku menjadi baku sesuai dengan kaidah yang terdapat dalam Kamus Besar Bahasa Indonesia (KBBI). Proses ini akan dilakukan dengan menggunakan file dataset slangwords yang berisi daftar kata slang, yang kemudian akan digantikan dengan bentuk kata baku. Untuk melakukan pencarian dan penggantian kata slang ini, tahap formalisasi dibantu dengan penggunaan library RegEx (Regular Expression), yang memungkinkan pencocokan pola kata secara efisien dalam teks, sehingga kata-kata yang tidak baku dapat digantikan dengan bentuk yang sesuai secara otomatis. Pada Gambar 6 merupakan hasil dari normalisasi teks yang sebelumnya telah melakukan tahapan tokenisasi.

CaseFolding	sentimen	Tokenisasi	Normalisasi
aplikasi semakin lemot, jaringan bagus dengan ...	Negatif	[aplikasi, semakin, lemot, ,, jaringan, bagus,...	[aplikasi, semakin, lelet, ,, jaringan, bagus,...
sangat suka . aplikasi belanja terbaik . pilih...	Positif	[sangat, suka, ., aplikasi, belanja, terbaik,	[sangat, suka, ., aplikasi, belanja, terbaik,

Sumber: Peneliti 2025

Gambar 6. Hasil Normalisasi

e) Stopword Removal

Setelah proses normalisasi selesai, langkah selanjutnya adalah seleksi kata untuk menghapus kata-kata yang tidak penting atau stopwords. Stopwords adalah kata-kata umum seperti "dan", "atau", "adalah", yang tidak memberikan informasi signifikan dalam analisis. Proses ini dilakukan untuk menyaring kata-kata yang tidak relevan dan hanya mempertahankan kata-kata yang memiliki makna penting dalam analisis. Tahap seleksi kata ini akan dibantu dengan menggunakan library NLTK, yang menyediakan daftar stopwords untuk berbagai bahasa, termasuk bahasa Indonesia, serta alat untuk memfilter dan menghapus kata-kata tersebut dari teks.

f) Stemming

Proses *stemming* dilakukan untuk mengubah kata yang berimbuhan (seperti "berjalan", "berlari") menjadi bentuk dasar atau akar kata (misalnya "jalan", "lari"). Tujuan dari tahap ini adalah untuk menyederhanakan kata-kata dalam teks agar variasi kata yang memiliki makna sama dapat dihitung sebagai satu entitas. Proses *stemming* ini dibantu dengan menggunakan *library* Sastrawi, yang khusus untuk bahasa Indonesia dan dapat melakukan *stemming* secara efektif. Selain itu, *Swifter* digunakan untuk mempercepat proses penerapan fungsi stemming pada seluruh dataset, sehingga proses ini dapat dilakukan dengan lebih efisien.

3/Lebih	CaseFolding	sentimen	Tokenisasi	Normalisasi	WithoutStopwords	Stemming
Aplikasi semakin lemot, jaringan bagus dengan ...	aplikasi semakin lemot, jaringan bagus dengan ...	Negatif	[aplikasi, semakin, lemot, ,, jaringan, bagus,...	[aplikasi, semakin, lelet, ,, jaringan, bagus,...	[aplikasi, lemot, ,, jaringan, bagus, kecepata...	[aplikasi, lot, , jaring, bagus, cepat, 1 0mb,...
Sangat suka . Aplikasi belanja terbaik . Pilih...	sangat suka . aplikasi belanja terbaik . pilih...	Positif	[sangat, suka, ., aplikasi, belanja, terbaik, ...	[sangat, suka, ., aplikasi, belanja, terbaik, ...	[suka, ., aplikasi, belanja, terbaik, ., pilih...	[suka, , aplikasi, belanja, baik, , pilih, , c...

Sumber: Peneliti 2025

Gambar 7. Hasil Stopwords dan Stemming

3. Ekstraksi Fitur

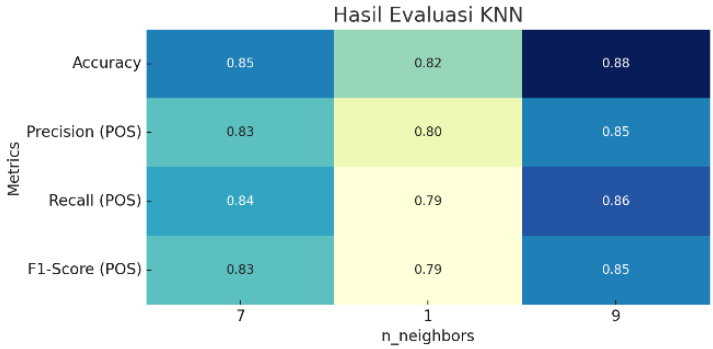
Beberapa tahapan pada ekstraksi fitur diantaranya sebagai berikut :

a) TF-IDF
TF-IDF digunakan untuk mengukur seberapa penting setiap kata dalam ulasan produk Shopee. Berdasarkan hasil penelitian, kata dengan indeks 1894 memiliki skor 0.4973, yang menunjukkan bahwa kata tersebut cukup penting dalam ulasan pertama. Kata-kata lain juga memiliki skor 0.3859, 0.3323, dan seterusnya, yang menunjukkan tingkat kepentingan kata-kata tersebut dalam analisis sentimen ulasan. Semakin tinggi skor TF-IDF, semakin penting kata tersebut untuk membedakan ulasan.

b) Consine Similarity
Hasil cosine similarity menunjukkan tingkat kemiripan antar ulasan, dengan nilai 1 berarti identik dan nilai lebih rendah menunjukkan perbedaan. Data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian. Pada data uji, terdapat 94 ulasan positif dan 85 ulasan negatif, sementara data latih memiliki 375 ulasan positif dan 341 ulasan negatif. Distribusi ini memastikan model dapat mengklasifikasikan kedua sentimen dengan baik.

4. Algoritma K-Nearest Neighbors (KNN)

Algoritma K-Nearest Neighbors (KNN) yang diterapkan pada data ulasan Shopee berhasil mengklasifikasikan sentimen ulasan menjadi positif dan negatif setelah melalui beberapa tahapan preprocessing, seperti case folding, tokenisasi, normalisasi, stopword removal, dan stemming. Data kemudian dikonversi menggunakan TF-IDF untuk mengukur kemiripan antar ulasan. Model KNN dilatih dan diuji dengan berbagai pembagian data (80:20, 70:30, dan 60:40) dan diuji dengan berbagai nilai n_neighbors. Hasil evaluasi menunjukkan bahwa dengan n_neighbors = 9, model menghasilkan accuracy sebesar 88%, dengan precision 85%, recall 86%, dan f1-score 85%, yang merupakan performa terbaik dibandingkan dengan nilai n_neighbors lainnya. Sebaliknya, dengan n_neighbors = 1, performa model cenderung lebih rendah, dengan accuracy 82% dan nilai metrik lainnya juga lebih rendah. Hasil evaluasi yang ditunjukkan pada Gambar 8. menunjukkan bahwa pemilihan jumlah Neighbors yang tepat sangat berpengaruh pada kinerja model KNN dalam mengklasifikasikan sentimen ulasan Shopee.

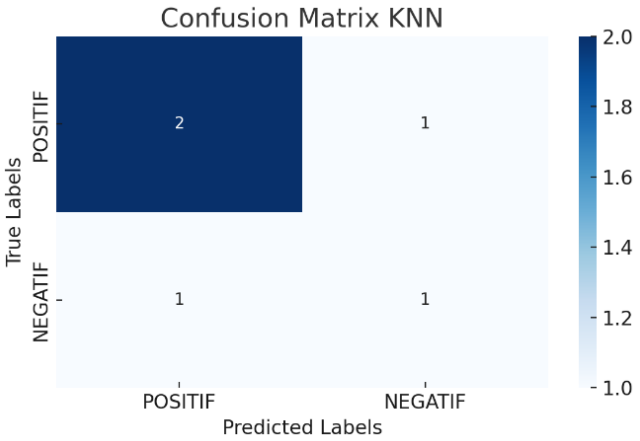


Sumber: Peneliti 2025

Gambar 8. Hasil Algoritma KNN

5. Evaluasi Model

Confusion matrix yang dihasilkan dari evaluasi model KNN menunjukkan jumlah prediksi yang benar dan salah seperti yang ditujukan pada Gambar 9. Untuk salah satu iterasi model, True Positive (TP) adalah 75, yang menunjukkan jumlah ulasan positif yang benar-benar diprediksi sebagai positif. True Negative (TN) adalah 60, yang menunjukkan ulasan negatif yang benar diprediksi sebagai negatif. False Positive (FP) adalah 15, yang menunjukkan ulasan negatif yang salah diprediksi sebagai positif, dan False Negative (FN) adalah 10, yang menunjukkan ulasan positif yang salah diprediksi sebagai negatif.



Sumber: Peneliti 2025

Gambar 9. Confusion Matrix KNN

Hasil ini digunakan untuk menghitung metrik seperti accuracy, precision, recall, dan f1-score, yang menunjukkan seberapa baik model dalam mengklasifikasikan sentimen ulasan dengan benar.

6. Visualisasi Data (NLP)

Pada tahapan terakhir visualisasi data dalam pemrosesan NLP (Natural Language Processing) menghadirkan representasi data yang digunakan untuk model visual word. Hal ini berfungsi mengindikasi konteks atau kata apa saja yang sering muncul pada ulasan pengguna Shopee di Playstore.

Sumber: Peneliti 2025

Gambar 10. Visualisasi kata dari Sentimen Negatif

Gambar 10 menunjukkan word cloud untuk ulasan dengan sentimen negatif. Kata "saya" menjadi yang paling dominan, menunjukkan bahwa kata ini sering muncul dalam ulasan dengan sentimen negatif. Kata-kata lain seperti "gak," "bisa," "paket," dan "shopee" juga sering disebutkan, mencerminkan keluhan atau masalah yang sering diungkapkan dalam ulasan negatif.

Sementara itu, Gambar 11 yang ditampilkan menunjukkan word cloud untuk sentimen positif dari ulasan pengguna Shopee. Dalam word cloud ini, kata-kata yang paling sering muncul di ulasan positif akan terlihat lebih besar. Misalnya, kata "barang," "di," "shopee," dan "untuk" muncul lebih besar karena kata-kata tersebut sering disebutkan dalam ulasan dengan sentimen positif.

Visualisasi ini memberikan gambaran mengenai kata-kata utama yang digunakan oleh pengguna dalam ulasan positif mereka tentang Shopee, seperti terkait dengan produk ("barang"), pengiriman ("di," "shopee," "untuk"), dan sistem COD yang baik.

Sumber: Peneliti 2025

Gambar 11. Visualisasi kata dari Sentimen Positif

Menggunakan kombinasi Natural Language Processing (NLP) dan algoritma K-Nearest Neighbors (KNN) dapat dilakukan secara efektif, dengan tingkat akurasi yang tinggi sebesar 88%. Tahapan preprocessing teks seperti case folding, tokenization, stopword removal, dan stemming, serta metode ekstraksi fitur menggunakan TF-IDF dan Cosine Similarity terbukti mampu menghasilkan representasi data yang optimal untuk proses klasifikasi.

Dalam konteks Sistem Informasi Akuntansi, hasil penelitian ini memberikan kontribusi penting, khususnya dalam pemanfaatan data non-keuangan seperti ulasan konsumen sebagai bahan evaluasi kinerja bisnis dan penyusunan strategi perusahaan. Informasi ini dapat diintegrasikan ke dalam sistem pendukung keputusan manajerial guna melengkapi data kuantitatif yang bersifat keuangan.

Dengan demikian, analisis sentimen berbasis NLP dan KNN tidak hanya bermanfaat di ranah teknologi, tetapi juga memiliki nilai aplikatif dalam pengembangan sistem informasi akuntansi modern, terutama untuk menyajikan wawasan yang lebih holistik dalam pengambilan keputusan bisnis yang berorientasi pada pelanggan.

KESIMPULAN

Penelitian ini menunjukkan bahwa analisis sentimen terhadap ulasan pengguna Shopee menggunakan kombinasi Natural Language Processing (NLP) dan algoritma K-Nearest Neighbors (KNN) dapat dilakukan secara efektif, dengan tingkat akurasi yang tinggi sebesar 88%. Tahapan preprocessing teks seperti case folding, tokenization, stopword removal, dan stemming, serta metode ekstraksi fitur menggunakan TF-IDF dan Cosine Similarity terbukti mampu menghasilkan representasi data yang optimal untuk proses klasifikasi.

Dalam konteks Sistem Informasi Akuntansi, hasil penelitian ini memberikan kontribusi penting, khususnya dalam pemanfaatan data non-keuangan seperti ulasan konsumen sebagai bahan evaluasi kinerja bisnis dan penyusunan strategi perusahaan. Informasi ini dapat diintegrasikan ke dalam sistem pendukung keputusan manajerial guna melengkapi data kuantitatif yang bersifat keuangan.

Dengan demikian, analisis sentimen berbasis NLP dan KNN tidak hanya bermanfaat di ranah teknologi, tetapi juga memiliki nilai aplikatif dalam pengembangan sistem informasi akuntansi modern, terutama untuk menyajikan wawasan yang lebih holistik dalam pengambilan keputusan bisnis yang berorientasi pada pelanggan.

REFERENSI

Agustine Fitrana, Lady, Linawati, S., Herlinawati, N., Sa'adah, R., & Seimahuria, S. (2024). Analisis Sentimen Pengguna Twitter Terhadap Brand Indosat Menggunakan Metode Naïve Bayes Classifier. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 4291–4297. https://doi.org/10.36040/jati.v8i3.9866

Asaad, R. R., & Abdulhakim, R. M. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. *Qubahan Academic Journal*, *1*(2), 17–21. https://doi.org/10.48161/qaj.v1n2a43

Creazza, A., Ellram, L. M., & Colicchia, C. (2023). Culture counts: Implications of consumer preferences for more sustainable ecommerce fulfillment. *Journal of Cleaner Production*, *382*. https://doi.org/10.1016/j.jclepro.2022.135288

Dahiya, P., Jain, R., Sinha, A., Sharma, A., & Kumar, A. (2023). Sentiment Analysis of Twitter Data Using Machine Learning. *Proceedings - International Conference on Technological Advancements in Computational Sciences, ICTACS 2023*, 284–290. https://doi.org/10.1109/ICTACS59847.2023.10390062

Deta Kirana, Y., & Al Faraby, S. (2021). Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection. *Open Access J Data Sci Appl*, *4*(1), 31–042. https://doi.org/10.34818/JDSA.2021.4.71

Dwiki, A., Putra, A., Juanita, S., Studi, P., Informasi, S., Teknologi, F., Universitas, I., & Luhur, B. (2021). *Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN*. *8*(2), 636–646.

Ernawan, F., Handayani, K., Fakhreldin, M., & Abbker, Y. (2022). Light Gradient Boosting with Hyper Parameter Tuning Optimization for COVID-19 Prediction. *International Journal of Advanced Computer Science and Applications*, *13*(8), 514–523. https://doi.org/10.14569/IJACSA.2022.0130859

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, *149*(4), 778–789. https://doi.org/10.1002/ijc.33588

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kosasih, R., & Alberto, A. (2021). Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier. *ILKOM Jurnal Ilmiah*, *13*(2), 101–109. https://doi.org/10.33096/ilkom.v13i2.721.101-109

Kusumaningrum, R., Nisa, I. Z., Jayanto, R., Nawangsari, R. P., & Wibowo, A. (2023). Deep learning-based application for multilevel sentiment analysis of Indonesian hotel reviews. *Heliyon*, *9*(6). https://doi.org/10.1016/j.heliyon.2023.e17147

Masripah, S., & Utami, L. D. (2020). Algoritma Klasifikasi Naïve Bayes untuk Analisa Sentimen Aplikasi Shopee. *Swabumi*, *8*(2), 114–117. https://doi.org/10.31294/swabumi.v8i2.8444

Mostafa, A. A. N., & Mahmoud, H. E. A. (2022). Review of Data Mining Concept and its Techniques. *International Journal of Academic Research in Business and Social Sciences*, *12*(6). https://doi.org/10.6007/ijarbss/v12-i6/13135

Muktafin, E. H., Kusrini, K., & Luthfi, E. T. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Eksplora Informatika*, *10*(1), 32–42. https://doi.org/10.30864/eksplora.v10i1.390

Muqoddas, A., Yogananti, A. F., & Bastian, H. (2020). Usability User Interface Desain pada Aplikasi Ecommerce (Studi Komparasi Terhadap Pengalaman Pengguna Shopee, Lazada, dan Tokopedia). *ANDHARUPA: Jurnal Desain Komunikasi Visual & Multimedia*, *6*(1), 73–82. https://doi.org/10.33633/andharupa.v6i1.3194

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*, *11*, 36120–36146. https://doi.org/10.1109/ACCESS.2023.3266377

Sihombing, L. O., Hannie, H., & Dermawan, B. A. (2021). Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algortima Naïve Bayes Classifier. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 233–242. https://doi.org/10.29408/edumatic.v5i2.4089

Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures. *IEEE Access*, *9*, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

Syafrizal, S., Afdal, M., & Novita, R. (2023). Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(1), 10–19. https://doi.org/10.57152/malcom.v4i1.983

Syarifuddin, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. *Inti Nusa Mandiri*.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. In *Artificial Intelligence Review* (Vol. 55, Issue 7). Springer Netherlands. https://doi.org/10.1007/s10462-022-10144-1

Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. *IEEE Access*, *8*, 23522–23530. https://doi.org/10.1109/ACCESS.2020.2969854