INCODING: Journal of Informatic and Computer Science Engineering
https://journal.mahesacenter.org/index.php/incoding/index || ISSN xxxx-xxxx (online)
Vol (Issue) Year: Page

DOI: **Genesis:**

Analisis Sentimen Produk Berdasarkan Review Pelanggan Shopee Menggunakan KNN

Sentiment Analysis of Shopee Product Reviews Using the K-Nearest Neighbors (KNN) Algorithm

Fira Irwannia[1)*], Andre Hasudungan Lubis[2)]
1)Prodi Teknik Informatika, Fakultas Teknik, Universitas Medan Area, Indonesia

Abstrak

Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap ulasan pelanggan mengenai produk mukena yang tersedia di aplikasi Shopee menggunakan algoritma *K-Nearest Neighbors* (KNN). Data yang digunakan merupakan data primer sebanyak 200 ulasan yang dikumpulkan secara manual. Proses analisis dimulai dari *preprocessing data* berupa *case folding*, tokenisasi, penghapusan *stopwords*, dan *stemming*, kemudian dilanjutkan dengan ekstraksi fitur menggunakan metode TF-IDF, dan klasifikasi menggunakan algoritma KNN. Penelitian ini juga melakukan evaluasi terhadap performa model dengan akurasi. Hasil pengujian menunjukkan bahwa proporsi *data training* dan nilai parameter *n_neighbors* sangat mempengaruhi akurasi model. Proporsi *data training* sebesar 90% dan *testing* 10% menghasilkan akurasi tertinggi sebesar 90%. Namun, ketika nilai *n_neighbors* = 3, proporsi 70:30 menghasilkan performa terbaik sebesar 81,67%. Penelitian ini menunjukkan bahwa KNN mampu digunakan sebagai metode yang efektif dalam analisis sentimen terhadap ulasan produk.
Kata Kunci: Analisis Sentimen, Ulasan Produk, KNN, Shopee, TF-IDF

Abstract

This study aims to conduct sentiment analysis on customer reviews of mukena products available on the Shopee application using the K-Nearest Neighbors (KNN) algorithm. The data used is primary data consisting of 200 reviews collected manually. The analysis process begins with data preprocessing such as case folding, tokenization, stopword removal, and stemming, followed by feature extraction using the TF-IDF method, and classification using the KNN algorithm. The model's performance is evaluated using a confusion matrix. The results show that the proportion of training data and the n_neighbors parameter significantly affect the model's accuracy. A 90% training and 10% testing proportion produced the highest accuracy of 90%. However, with n_neighbors = 3, the best performance was achieved with a 70:30 data split, reaching 81.67% accuracy. This study demonstrates that KNN is an effective method for sentiment analysis on product reviews.

Keywords: Sentiment Analysis, Product Reviews, KNN, Shopee, TF-IDF

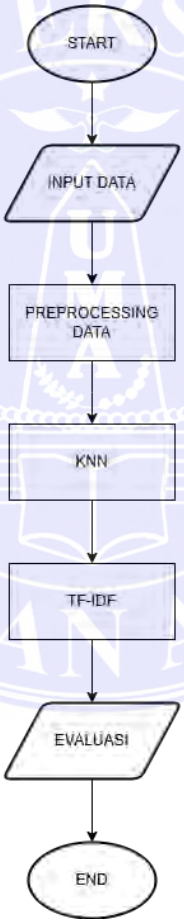
PENDAHULUAN

Dalam beberapa tahun terakhir, belanja *online* melalui platform *e-commerce* telah menjadi bagian yang tidak terpisahkan dari kehidupan masyarakat modern. Kemudahan akses, efisiensi waktu, dan berbagai pilihan produk yang ditawarkan menjadi alasan utama meningkatnya popularitas platform *e-commerce* di Indonesia. Salah satu platform yang berhasil menarik perhatian masyarakat adalah *Shopee*, dengan jumlah pengguna aktif bulanan mencapai 117 juta pada kuartal pertama tahun 2023 (Katadata, 2021). *Shopee* menempati posisi kedua sebagai situs *e-commerce* dengan jumlah kunjungan bulanan terbanyak di Indonesia pada kuartal pertama tahun 2021 (Cahyaningtyas dkk., 2021). Pengguna dapat menilai dan mengulas aplikasi di *Google Playstore* (Muttaqin & Kharisudin, 2021). Ulasan produk online mengurangi kemungkinan inkonsistensi pembelian dan membantu calon pembeli dalam mengambil keputusan (Alfaris & Kusnawi, 2023). *Playstore* merupakan salah satu layanan milik *Google* yang menyediakan berbagai jenis konten digital, seperti game, aplikasi, film, musik, dan buku (Supeli & Setiaji, 2023). Setiap aplikasi memiliki keunggulan dan kelemahan tersendiri yang dapat memicu tanggapan berbeda dari para pengguna, seperti rasa puas maupun tidak puas (Giovani dkk., 2020). Analisis sentimen terhadap ulasan aplikasi media sosial menjadi sarana yang bermanfaat bagi bisnis untuk memahami kebutuhan serta preferensi pelanggan, sekaligus mengidentifikasi permasalahan yang berpotensi merusak citra mereka (Ardiansyah dkk., 2023). Analisis sentimen terhadap produk juga merupakan metode yang efisien untuk mengetahui bagaimana pelanggan menilai suatu produk. Penelitian ini menunjukkan bahwa ulasan dari pengguna memiliki pengaruh terhadap minat beli dan menjadi faktor penting dalam peningkatan penjualan (Nurul & Isnalita, 2019). Ulasan produk pada platform *marketplace* memberikan informasi yang berguna bagi calon pembeli dalam mengambil keputusan pembelian dan mengurangi risiko ketidaksesuaian produk. *Marketplace* yang dijadikan objek kajian dalam penelitian ini adalah *Shopee*, yang merupakan salah satu platform *e-commerce* populer di Indonesia maupun mancanegara, karena menawarkan kemudahan dalam transaksi jual beli berbagai kebutuhan sehari-hari (Ruger dkk., 2021). Menurut (Lubis dkk., 2019), perilaku belanja online milenial dipengaruhi oleh kepercayaan, sikap, norma subjektif, dan kontrol perilaku yang dirasakan.

INCODING: Journal of Informatic and Computer Science Engineering *ISSN 2622-3740 (Online)*
Vol 1, No. 1, Agustus 2018: 1 -10

Melakukan analisis sentimen secara manual merupakan pekerjaan yang menantang karena melibatkan volume data yang besar, adanya ulasan pelanggan yang tidak seragam, serta keterbatasan kemampuan manusia dalam melakukan analisis secara manual. Oleh karena itu, penelitian ini mengusulkan solusi dengan menggunakan metode *K-Nearest Neighbors* (K-NN) untuk mengukur tingkat akurasi dalam mengklasifikasikan sentimen produk di platform *Shopee* (Keaan dkk., 2019). Algoritma *K-Nearest Neighbors* merupakan algoritma popular yang termasuk dalam grup *instance-based learning*. K-Nearest Neighbors sendiri merupakan metode *lazy-learning* yang melakukan klasifikasi berdasarkan kedekatan jarak antara data yang dianalisis (Puspita & Widodo, 2021).

METODE PENELITIAN



Gambar 1 Metode Penelitian

Penelitian ini menerapkan algoritma KNN untuk melakukan analisis sentimen terhadap komentar produk mukena. Metode penelitian ini bertujuan untuk memberikan gambaran yang jelas dan sistematis mengenai tahapan-tahapan utama, mulai dari input

data komentar, praproses data, penerapan metode TF-IDF, dan penggunaan algoritma KNN, hingga evaluasi dalam bentuk akurasi.

2.1 Pengumpulan Data

Penelitian ini menggunakan data primer yang diperoleh langsung oleh penulis melalui proses pengumpulan secara manual. *Dataset* yang berhasil dihimpun terdiri atas 200 ulasan produk mukena yang tersedia di aplikasi *Shopee*. Proses pengumpulan ini dilakukan untuk memperoleh data ulasan yang autentik dan relevan, yang akan dijadikan dasar analisis dalam penelitian ini. Maka contoh data yang sudah berhasil dikumpulkan dapat dilihat pada Tabel 1.

Tabel 1 Contoh Dataset dari Shopee

No	Nama Pengguna	Komentar
1.	m1laa	pertama kali order alhamdulillah langsung puas, pengirimian jg cepat, packing mantap
2.	Ujibon	Kainnya lembut dan sesuai gambar. Mukena sangat nyaman dan akan membeli kembali
3.	ephy_ephy	Bahan adem sptnya, penjual ramah, cm warna almond milknya mirip grey ya?
4.	esampalupi	Bahannya halus, rendanya ga lebay. Untuk hadiah mama dan ipar semoga sukaa Terima kasih vouchernya
5.	dewi8150	Bahannya bagus adem nyaman dipakai sangat cocok, cuma warna tidak sama dengan dfoto dan asli Warna asli lebih ke coklat muda ato mocca

2.2 *Preprocessing Data*

Preprocessing data merupakan tahap awal dalam pengolahan data yang bertujuan untuk mempermudah proses analisis dan pemrosesan lebih lanjut. Tahapan ini dilakukan untuk membersihkan data dari gangguan (*noise*), menyederhanakan dimensi data, serta menyusun data agar lebih terstruktur (Darmawan dkk., 2023). Beberapa teknik yang digunakan dalam preprocessing data antara lain: *case folding*, yang berfungsi

INCODING: Journal of Informatic and Computer Science Engineering *ISSN 2622-3740 (Online)*
Vol 1, No. 1, Agustus 2018: 1 -10

menyamakan semua karakter, sehingga kata seperti "The" dan "THE" dianggap identik; *stopword removal*, yang berguna untuk menghapus kata-kata yang dianggap tidak penting seperti kata hubung, biasanya menggunakan pendekatan *bag of words*; serta *stemming*, yaitu proses mengubah kata yang memiliki imbuhan menjadi bentuk dasarnya (Syam dkk., 2024). Pada penelitian ini, proses *preprocessing* meliputi beberapa tahapan, yaitu *case folding* untuk menyamakan format huruf, *tokenization* untuk Memecah teks menjadi unit-unit yang lebih kecil. Penghapusan *stopwords* yang tidak memberikan makna signifikan, serta *stemming* untuk mereduksi kata ke bentuk dasarnya. Hasil dari *preprocessing* data dapat dilihat pada Tabel 2.

Tabel 2 Hasil *Preprocessing Data*

No	Komentar
1.	['pertama', 'order', 'alhamdulillah', 'langsung', 'puas', 'kirim', 'packing', 'mantap']
2.	['kain', 'lembut', 'sesuai', 'gambar', 'mukena', 'sangat', 'nyaman', 'beli', 'kembali']
3.	['bahan', 'adem', 'penjual', 'ramah', 'warna', 'almond', 'milk', 'mirip', 'grey']
4.	['bahan', 'halus', 'renda', 'lebay', 'hadiah', 'mama', 'ipar', 'semoga', 'suka', 'terima', 'kasih', 'voucher']
5.	['bahan', 'bagus', 'adem', 'nyaman', 'pakai', 'cocok', 'warna', 'foto', 'asli', 'warna', 'asli', 'coklat', 'muda']

2.3 TF-IDF

Setelah tahap *preprocessing* komentar selesai dilakukan, langkah berikutnya adalah memisahkan elemen kata dengan menggunakan teknik *Term Frequency-Inverse Document Frequency* (TF-IDF). TF-IDF merupakan metode pembobotan yang menghitung frekuensi kemunculan kata dalam satu dokumen (TF) serta frekuensi kata tersebut dalam keseluruhan kumpulan dokumen (IDF). Tujuan dari pendekatan ini adalah untuk menyeimbangkan perbedaan panjang dokumen sehingga menghasilkan ukuran bobot

yang lebih akurat dan dapat diandalkan (Astiningrum, Batubulan, & Sias, 2020). TF-IDF sendiri merupakan kombinasi dari dua konsep utama, yaitu *term frequency* (TF) dan *inverse document frequency* (Rozi dkk., 2021):

$$TF = \frac{t}{d} \tag{1}$$

$$idf = \log\left(\frac{N}{df(t)}\right) \tag{2}$$

$$TFidf = TF.idf \tag{3}$$

Keterangan:
t = jumlah kemunculan kata tertentu dalam dokumen d
d = total keseluruhan kata pada dokumen
N = total dokumen yang ada
df(t) = jumlah dokumen yang memiliki kata t

2.4 *K-Nearest Neighbors* (KNN)

Algoritma *K-Nearest Neighbors* (KNN) menjadi salah satu metode populer dalam machine learning karena kemudahan implementasinya dalam menangani permasalahan yang kompleks. KNN juga efektif dalam mengatasi ketidakseimbangan kelas pada *dataset*, karena proses klasifikasinya didasarkan pada mayoritas tetangga terdekat. Dengan pendekatan ini, KNN tetap mampu menghasilkan prediksi yang baik meskipun terdapat kelas minoritas yang kurang dominan. Selain untuk klasifikasi, algoritma ini juga dapat diterapkan pada masalah regresi, di mana nilai prediksi dihitung dari rata-rata nilai tetangga terdekat. Berikut ini adalah tahapan-tahapan dalam algoritma KNN (Alfaris & Kusnawi, 2023):

1) Menentukan nilai K sebagai parameter utama.
2) Menghitung jarak antara data uji dan setiap data pada dataset pelatihan, menggunakan metrik tertentu jika atribut bersifat numerik

INCODING: Journal of Informatic and Computer Science Engineering *ISSN 2622-3740 (Online)*
Vol 1, No. 1, Agustus 2018: 1 -10

$$D(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Keterangan:

D (xi, yi) = jarak

Xi = data latih

Yi = data uji

i = variabel data

n = Dimensi data

3.) Mengurutkan jarak yang telah dihitung secara menurun (descending)

4.) Memilih K data dengan jarak terdekat dari data uji.

5.) Mnentukan kelas mayoritas dari K tetangga tersebut dan mengklasifikasikan data uji ke dalam kelas tersebut.

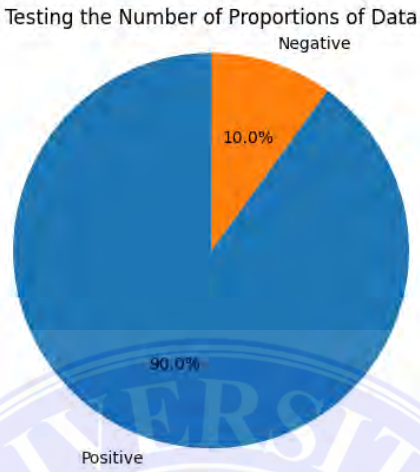
HASIL DAN PEMBAHASAN

Tabel 3 Pengujian Jumlah Proporsi Data

Pengujian	*Data Training*	*Data Testing*	Performa
1	70 %	30 %	73,33 %
2	80 %	20 %	82,5 %
3	90 %	10 %	90 %

Pada Tabel 3 Pengujian 1: Model dilatih dengan 70% data dan diuji dengan 30% data, menghasilkan performa akurasi sebesar 73,33%. Pengujian 2: Ketika proporsi data pelatihan ditingkatkan menjadi 80% dan data pengujian menjadi 20%, performa model naik menjadi 82,5%. Pengujian 3: Dengan proporsi data pelatihan tertinggi (90%) dan data pengujian paling sedikit (10%), model mencapai akurasi tertinggi sebesar 90%. Dari tabel ini terlihat bahwa semakin besar proporsi *data training* yang digunakan, semakin baik performa model. Hal ini menunjukkan bahwa model memiliki kemampuan belajar

yang lebih baik saat diberikan lebih banyak data untuk pelatihan. Namun, penting juga untuk menjaga keseimbangan agar data testing cukup representatif untuk mengukur generalisasi model.



Gambar 2 Hasil Sentimen Berdasarkan Jumlah Proporsi Data

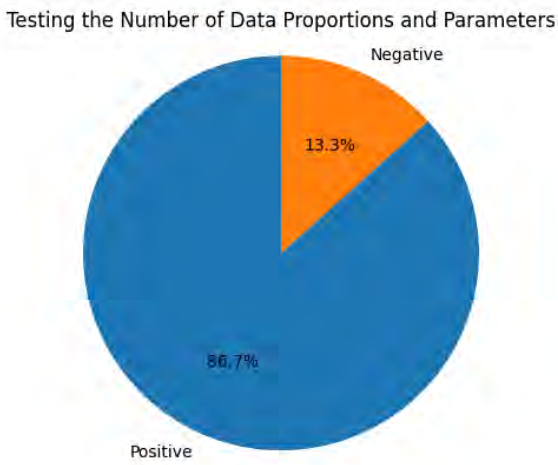
Berdasarkan Gambar 2 dapat dilihat bahwa sentimen positif terhadap produk mukena mendominasi yaitu positif sebesar 90 % dan negatif 10 %.

Tabel 4 Pengujian Jumlah Proporsi Data & Parameter

Pengujian	*Data Training*	*Data Testing*	*N_neighbors*	Performa
1	70 %	30 %	3	81,67 %
2	80 %	20 %	3	75 %
3	90 %	10 %	3	70 %

Pada Tabel 4 Pengujian 1, dengan 70% *data training* dan 30% *data testing*, model menghasilkan performa terbaik sebesar 81,67%. Pengujian 2, proporsi *training* ditingkatkan menjadi 80%, namun performa justru menurun menjadi 75%. Pengujian 3, dengan 90% *data training* dan hanya 10% *data testing*, performa kembali menurun menjadi 70%. Pemilihan proporsi *data training* dan *testing* yang seimbang sangat penting. Dalam kasus ini, proporsi 70:30 memberikan performa terbaik untuk model KNN dengan *n_neighbors* = 3.

INCODING: Journal of Informatic and Computer Science Engineering *ISSN 2622-3740 (Online)*
Vol 1, No. 1, Agustus 2018: 1 -10



Gambar 3 Hasil Sentimen Berdasarkan Jumlah Proporsi Data & Parameter

Berdasarkan Gambar 3 dapat dilihat bahwa sentimen positif juga masih mendominasi yaitu positif sebesar 86,7 % dan negatif 13,3 %.

SIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Nearest Neighbors (KNN) dapat digunakan secara efektif untuk melakukan klasifikasi sentimen terhadap ulasan produk. Proporsi *data training* dan pemilihan nilai *n_neighbors* berpengaruh signifikan terhadap performa model. Pengujian menunjukkan bahwa proporsi data 90:10 memberikan akurasi tertinggi, sementara nilai *n_neighbors* = 3 dengan proporsi data 70:30 menghasilkan performa terbaik. Adapun hasil sentimen positif sebesar 86,7 % sedangkan sentimen negatif sebesar 13,3%. Oleh karena itu, pemilihan parameter yang tepat sangat penting dalam penerapan algoritma KNN untuk analisis sentimen.

DAFTAR PUSTAKA

Alfaris, S., & Kusnawi. (2023). Komparasi Metode KNN dan Naive Bayes Terhadap Analisis Sentimen Pengguna Aplikasi Shopee. *Indonesian Journal of Computer Science*, 2766–2776.

Ardiansyah, D., Saepudin, A., Aryanti, R., Fitriani, E., & Royadi. (2023). Analisis Sentimen Review Pada Aplikasi Media Sosial Tiktok Menggunakan Algoritma K-NN dan SVM Berbasi PSO. *Jurnal Informatika Kaputama*, 233–241.

Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *Jurnal Teknologi Informasi*, 174–184.

Darmawan, M. B. A., Dewanta, F., & Astuti, S. (2023). Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naive Bayes untuk Prediksi Banjir di Desa Dayeuhkolot. *TELKA*, 52–61.

Giovani, A. P. A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru di Twitter Menggunakan Algoritma Klasifikasi. *Jurnal TEKNOINFO*, 116–124.

Katadata. (2021). *Market Size of Indonesia's Hijab Industry Reaches Rp 33 Trillion*.

Keaan, L. S. W. W., Indirati., & Marji. (2019). Analisis Sentimen Review Shopee Berbahasa Indonesia Menggunakan Improved K-Nearest Neighbor dan Jaro Winkler Distance. *J. Pengemb. Tecnol. Inf. dan Ilmu komput*, 7172–7179.

Lubis, A. H., Amelia, W. R., Ramadhani, S. N., Pane, A. A., & Aryza, S. (2019). Indonesian millennials' behavior intention to online shopping through instagram. *International Journal of Scientific and Technology Research*, *8*(11), 2466–2471.

Muttaqin, M. N., & Kharisudin, I. (2021). Analisis Sentimen Aplikasi Gojek Menggunakan Support Vector Machine Dan K-Nearest Neighbor. *Jurnal of Mathematics*, 22–27.

Nurul, M., & Isnalita, I. (2019). Pengaruh Jumlah Pengunjung, Ulasan Produk, Reputasi Toko Dan Status Gold Badge Pada Penjualan Dalam Tokopedia. *E-Jurnal Akunt*, 1855–1865.

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naive Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *J. Inform. Univ. Pamulang*, 646.

Rozi, F., Sukmana, F., & Adami, M. N. (2021). Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency - Inverse

INCODING: Journal of Informatic and Computer Science Engineering *ISSN 2622-3740 (Online)*
Vol 1, No. 1, Agustus 2018: 1 -10

Document Frequency (TF-IDF). *Jurnal Informatika Merdeka Pasuruan*, 1–5.

Ruger, A. H., Suyanto, M., & Kurniawan, M. P. (2021). Sentiment Analisis Pelanggan Shopee di Twitter dengan Algoritma Naive Bayes. *J. Inf, Technol*, 26–29.

Supeli, M. F. F., & Setiaji. (2023). Klasifikasi Sentimen Positif dan Negatif Pada Aplikasi Vidio Dengan Algoritma K-Nearest Neighbor. *Infonesian Journal Computer Science*, 7–15.

Syam, A. A., M, G. H., Salim, A., Surianto, D. F., & B, M. F. (2024). Analisis Teknik Preprocessing Pada Sentimen Masyarakat Terkait Konflik Israel-Palestina Menggunakan Support Vector Machine. *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, 1464–1472.

INCODING: Journal of Informatics and Computer Science Engineering
ISSN 2776-432X (Online - Elektronik)
https://journal.mahesacenter.org/index.php/incoding

Date: April, 24th 2025

LETTER OF ACCEPTANCE
Paper Number #865

Dear, Fira Irwannia, Andre Hasudungan Lubis

This is to inform you that the manuscript entitled: **"Analisis Sentimen Produk Berdasarkan Review Pelanggan Shopee Menggunakan KNN",** which was sent on **April 19th 2025,** is **ACCEPTED.**

We keep to ensuring a high standard of articles published in the **INCODING: Journal of Informatics and Computer Science Engineering,** and the manuscript that is being sent to you has been submitted after a first selection process based on the agreement of the **Associate Editors.** In general, the standard of manuscripts forwarded to me after the vetting is **good.**

This paper is well organized and follows the manuscript guidelines of the journal to a large extent. The introduction section is good and shows the importance of the study. The literature review is adequate. The outcomes of the study are consistent with the findings. The approach used is praiseworthy. In my opinion, it should be published without **revision again**

Based on the review results, this manuscript is **ACCEPTED**, and **PUBLISHED** in **Mei 2025** for **Volume 3, No. 2, 2025.**

Thank you very much for your contribution. Congratulations on a wonderful job.

Warmest Regards,
Editor In Chief

INCODING
2776-432X (Online - Elektronik)

Agung Suharyanto, S.Sn, M.Si.

Editorial Office:
Mahesa Research Center
Perumahan Griya Nafisa 2, Blok A No 10, Jalan Benteng Hilir
Titi Sewa, RT 06, Dusun XVI Flamboyan,
Kecamatan Percut Sei Tuan, Deli Serdang, 20371
Sumatera Utara, Indonesia
Phone: 082364935277
Email: incodingjournal@gmail.com