

ANALISIS SENTIMEN ULASAN E-COMMERCE SHOPEE DENGAN MENGGUNAKAN ALGORITMA NAIVE BAYES

Jeny Angreyani¹, Yonky Pernando^{1*}

¹Teknik Informatika, Universitas Universal

**email*: yongkyfernando194@gmail.com

Abstract: In this study, an analysis of the use of the Naive Bayes algorithm for sentiment analysis of reviews from Shopee app users on the Google Play Store was conducted, with classification divided into three categories: positive, negative, and neutral. To improve data quality, a preprocessing process was carried out with stages of cleaning, case folding, normalization, stop word removal, stemming, and tokenizing. Next, the text is formatted using the TF-IDF method to facilitate classification. For this data, the Naive Bayes model is used, which has an accuracy rate of 87% in detecting sentiment. Positive and negative categories can be easily identified compared to neutral sentiments due to the smaller amount of neutral data. Overall, the Naive Bayes algorithm successfully analyzed user sentiments well. The research can be developed with other algorithm methods, such as SVM, K-NN, or Decision Tree, in order to compare the performance of various algorithms.

Keywords: sentiment analysis; naive bayes; user reviews; e-commerce; shopee

Abstrak: Dalam penelitian ini dilakukan analisis penggunaan algoritma Naive Bayes untuk analisis sentimen review dari pengguna aplikasi Shopee di Google Play Store, klasifikasi dibagi menjadi 3 kategori yaitu positif, negatif, dan netral. Untuk meningkatkan kualitas data, dilakukan proses preprocessing dengan tahap cleaning, case folding, normalisasi, stopword removal, stemming, dan tekonezing. Selanjutnya, teks diformat menggunakan metode TF-IDF untuk memudahkan klasifikasi. Untuk data ini, model Naive Bayes digunakan, yang memiliki tingkat akurasi 87% dalam mendeteksi sentimen. Kategori positif dan negatif dapat dengan mudah diidentifikasi dibandingkan sentimen netral karena jumlah data netral yang lebih sedikit. Secara keseluruhan, algoritma Naive Bayes berhasil menganalisis perasaan pengguna dengan baik. Penelitian dapat dikembangkan dengan algoritma metode lain, seperti SVM, K-NN, atau Decision Tree, guna membandingkan kinerja berbagai algoritma.

Kata kunci: analisis sentiment; naive bayes; ulasan pengguna; e-commerce; shopee

PENDAHULUAN

Shopee, yang pertama kali muncul di Indonesia pada tahun 2015, adalah salah satu platform e-commerce terbaik di Indonesia. Aplikasi Shopee, yang tersedia di Play Store, telah mencapai 50 juta pengguna dalam hanya beberapa tahun. Angka ini menunjukkan bahwa jumlah penjualan dan pembelian meningkat setiap hari. Kesuksesan ini dipengaruhi oleh berbagai fitur Shopee, seperti pilihan produk yang luas, harga yang kompetitif, dan proses pembayaran yang mudah [1].

Proses analisis sentimen adalah mencari kesamaan individu melalui kalimat atau teks, serta mengumpulkan data kata untuk mengetahui isu-isu sentimen yang sering

diungkapkan di media sosial, baik dalam bentuk positif, negatif, atau netral. Biasanya tahapan analisis ini otomatis dilakukan melalui berbagai perangkat lunak atau platform yang mendukung analisis teks [2].

Penilaian sentimen mempermudah transformasi data acak menjadi data terstruktur. Beberapa penelitian sebelumnya telah membahas penggunaan metode Naive Bayes untuk analisis sentimen. Peneliti memilih metode Naive Bayes karena strukturnya yang praktis, rumusnya sederhana, dan akurasinya lebih tinggi dibandingkan metode lainnya. Ciri utama dari algoritma ini adalah adanya asumsi yang sangat kuat (naif) terhadap independensi setiap kondisi atau kejadian [3]. Analisis data sentimen adalah metode untuk mengidentifikasi, mengekstraksi, atau menemukan informasi emosional dari teks, terutama dalam konteks opini atau emosi pengguna. Analisis ini terutama bertujuan untuk memahami perasaan atau pendapat yang diungkapkan pengguna melalui ulasan di Shopee [4]. Penelitian ini mengadopsi metode database Knowledge Discovery (KDD). Prosesnya meliputi beberapa tahapan yaitu seleksi data, preprocessing, transformasi, data mining dan evaluasi [5].

Dalam studi ini, kami akan memperkenalkan metode yang digunakan dalam penilaian sentimen, mengumpulkan data ulasan pengguna untuk aplikasi Shopee, mendidik dan menguji Naive Bayes, serta memahami data untuk penilaian sentimen guna mendapatkan wawasan penting. Penelitian ini diharapkan dapat memberikan kontribusi penting terhadap analisis sentimen pada aplikasi seluler dan memberikan pemahaman yang lebih mendalam mengenai persepsi pengguna terhadap aplikasi toko Shopee [6].

Dalam kontras dengan penelitian sebelumnya, penelitian ini memanfaatkan data komentar terbaru dan mengamati perbaikan aplikasi Shopee setelah penelitian sebelumnya. Dalam penelitian ini, 100.000 data poin dari komentar di Play Store akan digunakan untuk memastikan kecocokan teknik Naive Bayes dalam menganalisis sentimen pengguna terhadap aplikasi Shopee secara menyeluruh. Berbeda dengan studi sebelumnya, Studi ini akan meningkatkan analisis sentimen pengguna untuk aplikasi Shopee karena komentar pengguna terbagi menjadi tiga kategori: positif, negatif, dan netral [7].

Tujuan dari penelitian ini adalah untuk mengevaluasi pendeteksian ulasan di platform e-commerce Shopee, dengan menelaah pro dan kontra dari algoritma yang digunakan dalam penelitian deteksi ulasan e-commerce. Menurut penelitian Tania Puspa Rahayu Sanjaya et al., 2023, dalam analisis sentimen ulasan e-commerce Shopee menggunakan algoritma Naive Bayes dan SVM, mengidentifikasi bahwa dataset yang tidak seimbang atau mengandung noise dapat memengaruhi hasil klasifikasi. Begitu juga penelitian yang dilakukan oleh Ade et al., 2024 menemukan bahwa metode Naive Bayes berhasil mengklasifikasikan ulasan pelanggan Shopee Mall dengan tingkat akurasi yang bervariasi, namun menyarankan bahwa akurasi dapat ditingkatkan dengan penggunaan dataset yang lebih besar dan pemrosesan data yang lebih baik, seperti stemming dan lemmatization.

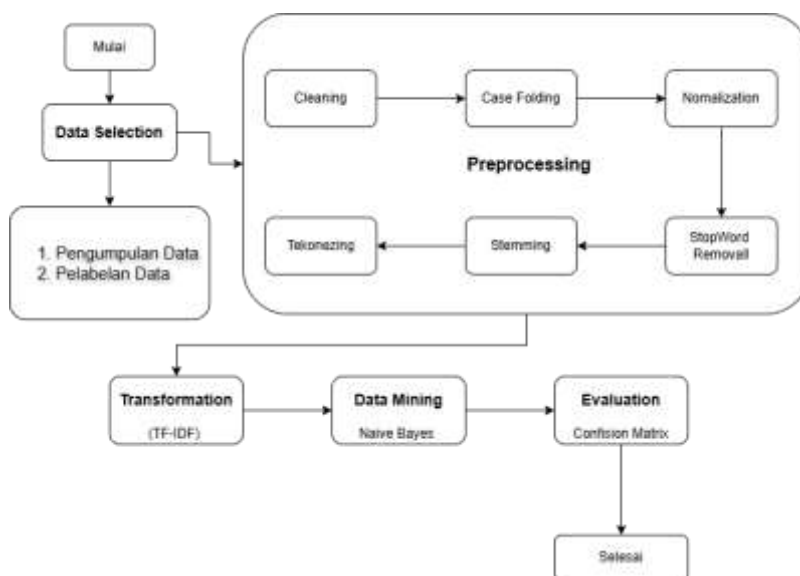
Tanggapan pengguna yang sudah menggunakan aplikasi dapat mempengaruhi keputusan menggunakan aplikasi tersebut, jadi review yang tidak baik bisa berdampak pada penjualan di platform itu. Berkat tinjauan dari pengguna sebelumnya, info mengenai kualitas barang dan jasa jadi lebih akurat, dan banyak pengguna online mempercayai saran serta pandangan tersebut. Pemodelan Naive Bayes akan dieksplorasi

untuk menilai performanya dalam mengidentifikasi ulasan. Meskipun demikian, kelemahan algoritma ini masih terdapat dalam melakukan deteksi ulasan secara tepat, yang menjadi fokus evaluasi dalam penelitian ini. Oleh karena itu Penelitian ini diharapkan dapat memberikan manfaat besar dalam memahami seberapa efektif algoritma naive bayes dalam mengidentifikasi ulasan di platform e-commerce seperti Shopee. Dengan mengevaluasi kelebihan dan kelemahan algoritma yang digunakan, penelitian ini diharapkan dapat memberikan saran yang dapat meningkatkan ketepatan deteksi ulasan untuk membantu platform dan penjual merespons umpan balik pelanggan dengan lebih efisien. Harapannya, peningkatan sistem deteksi ini akan meningkatkan kualitas layanan e-commerce dan memperbaiki kepuasan pengguna secara keseluruhan.

METODE

Metode Knowledge Discovery in Database digunakan dalam penelitian ini karena kelebihanannya sebagai proses terstruktur untuk mengidentifikasi pola dalam kumpulan data. Penemuan pengetahuan basis data (KDD) adalah penerapan metode ilmiah dalam penambangan data. Penambangan data (DM) adalah inti dari proses KDD dan melibatkan penarikan kesimpulan dari algoritma yang mengeksplorasi data, mengembangkan model, dan menemukan pola yang sebelumnya tidak diketahui. KDD mengidentifikasi informasi berharga dan pola tersembunyi dalam database besar yang sebelumnya belum ditemukan namun memiliki keuntungan penting.

Terdapat 5 proses text mining pada tahapan KDD yaitu seleksi data, preprocessing, transformasi, data mining dan evaluasi.



Gambar 1. Diagram Metode Knowledge Discovery in Database

Tahap pertama, seleksi data fase ini merupakan proses pemilihan atau penyiapan data yang akan dijadikan atribut, indikator, atau target pada dataset yang besar. Pada tahap awal ini akan dilakukan pengumpulan data dan pelabelan data. Data diperoleh dengan menggunakan teknik web scraping yang diprogram dengan Python.

Tahap kedua, preprocessing yaitu text preprocessing dilakukan untuk mempersiapkan teks sebelum pengujian atau pelatihan, bertujuan mengurangi noise dalam data sehingga dapat meningkatkan kinerja classifier dan mempercepat proses klasifikasi. Tahap ini terdapat langkah-langkah yang harus dilakukan, berikut adalah penjelasannya.

Langkah dimulai dari cleaning, data dibersihkan dengan menghilangkan fitur-fitur yang tidak berguna untuk proses klasifikasi, seperti huruf, emoji, URL, hashtag, dan simbol ataupun kata-kata yang ditulis lebih dari dua kali. Proses case folding adalah proses mengubah seluruh huruf menjadi huruf kecil. Pada proses ini karakter-karakter A-Z yang terdapat pada data diubah ke dalam karakter a-z. Tahap normalization adalah proses pemrosesan atau perbaikan kata-kata tidak baku yang disebabkan oleh kesalahan ketik, penyingkatan, dan lain sebagainya, untuk diubah menjadi kata-kata yang sesuai dengan bentuk baku. Stopword removal dilakukan untuk menghapus kata-kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak ada artinya. Contoh stopwords dalam bahasa Indonesia adalah “yang”, “dan”, “di”, “dari”, dan seterusnya. Ide di balik penggunaan kata-kata berhenti adalah dengan menghilangkan kata-kata yang mengandung informasi rendah dari teks sehingga kita dapat fokus pada kata-kata penting. Kemudian proses stemming digunakan untuk menggambarkan pemetaan dan penguraian kata. Kami telah menyiapkan pustaka Python Sastrawi yang dapat digunakan untuk melakukan stemming dalam bahasa Indonesia. Perpustakaan Cendekiawan menggunakan algoritma Nazief dan Adriani untuk stemming bahasa Indonesia. Tahap tokenizing merupakan proses pemisahan teks menjadi salah satu komponen yang disebut sebagai token untuk analisis, yang dikenal sebagai tokenisasi. Sebuah token dapat terdiri dari kata, simbol, angka, tanda baca, dan elemen penting lainnya. Tokenisasi dapat terjadi dalam sebuah paragraf atau kalimat, tetapi dalam proses penerjemahan bahasa alami, "kata" adalah definisi dari sebuah token.

Tahap ketiga, transformation adalah proses mengubah dokumen menjadi representasi yang memenuhi persyaratan analitis, sering disebut sebagai transformasi teks atau transformasi atribut. Dalam studi ini, pendekatan TF-IDF (Term Frequency-Inverse Document Frequency) digunakan karena memiliki potensi untuk menghasilkan vektor kata yang lebih mudah dipahami. Term Frequency (TF) adalah proses yang menghitung frekuensi kemunculan setiap kata dalam dataset, sementara Inverse Document Frequency (IDF) adalah proses yang menentukan jumlah dokumen yang mengandung kata tertentu. Pendekatan ini penting karena membantu mengurangi gangguan dalam data, sehingga membuat algoritma Naive Bayes lebih efisien dan meningkatkan akurasinya dalam analisis.

Tahap keempat, data mining adalah tahap di mana data yang telah dimodifikasi digunakan untuk melakukan proses klasifikasi dengan algoritma Naive Bayes. Teknik Naive Bayes yang diterapkan dalam model klasifikasi ini dikenal sebagai multinomial Naive Bayes. Berikut adalah persamaan yang digunakan dalam algoritma Naive Bayes untuk mendukung proses klasifikasi.

$$P(H|X) \quad (1)$$

$P(H|X)$ adalah probabilitas bahwa H adalah sebuah hipotesis berdasarkan X jika X adalah sebuah fakta dan H adalah sebuah hipotesis. Dengan kata lain, $P(H|X)$ adalah kemungkinan bahwa H muncul setelah X . $P(X|H)$ adalah kemungkinan bahwa X benar dengan syarat H , atau kemungkinan bahwa X benar jika H benar.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

Probabilitas awal dari hipotesis H dilambangkan dengan $P(H)$, sementara probabilitas awal dari hipotesis X dilambangkan dengan $P(X)$.

Tahap kelima, evaluasi merupakan langkah terakhir di mana model yang telah dikembangkan sebelumnya diuji untuk menentukan akurasi menggunakan algoritma Naive Bayes, termasuk akurasi dari kategori sentimen ketiga. Selain itu, *confusion matrix* digunakan untuk menilai kinerja model yang dikembangkan dengan mempertimbangkan beberapa metrik, yaitu accuracy, precision, recall, f1 score, dan support. Pendekatan ini memastikan bahwa model dapat diukur secara menyeluruh untuk mengevaluasi efektivitasnya.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini terdiri dari ulasan pengguna aplikasi e-commerce Shopee yang diambil dari Google Play Store. Data tersebut bersumber dari situs Kaggle dalam format *comma separated value* (.csv). Ulasan mencakup periode Juni 2024 hingga September 2024, dengan fokus pada komentar mengenai performa aplikasi Shopee. Dataset yang terkumpul berjumlah 100.000 ulasan, dengan empat kolom data. Untuk mempermudah analisis, proses pelabelan dilakukan berdasarkan rating bintang yang diberikan pengguna. Ulasan dengan skor bintang ≤ 2 dikategorikan sebagai sentimen negatif, skor 3 sebagai sentimen netral, dan skor ≥ 4 sebagai sentimen positif. Pendekatan ini bertujuan untuk mengelompokkan data sesuai polaritas sentimen.

Table 1. Keterangan Kolom Dataset

Nama Kolom	Deskripsi
Username	Berisi nama pengguna yang memberikan ulasan
Score	Berisi rating atau skor yang diberikan
at	Berisi waktu ulasan diberikan dalam format date-time
Content	Berisi teks ulasan yang diberikan

Tahap preprocessing merupakan langkah penting untuk mempersiapkan data yang telah terkumpul dan diberi label agar siap digunakan dalam analisis lebih lanjut. Proses ini terdiri atas enam langkah utama, yaitu cleaning, case folding, normalization,

stopword removal, stemming, dan tokenizing. Semua langkah dilakukan menggunakan Jupyter Notebook (Google Colab) dengan bahasa pemrograman Python. Langkah pertama adalah cleaning, yaitu membersihkan data dengan menghilangkan tanda baca, angka, karakter khusus, dan link URL untuk mengurangi kompleksitas saat analisis dilakukan. Langkah berikutnya adalah case folding, di mana semua huruf dalam teks diubah menjadi huruf kecil untuk menghindari perbedaan interpretasi antara huruf besar dan kecil. Setelah itu, data dinormalisasi dengan mengganti kata-kata tidak baku atau singkatan menjadi bentuk standar untuk menyamakan variasi kata yang memiliki makna serupa.

Proses stopwords removal dilakukan untuk menghapus kata-kata umum seperti “di,” “ke,” atau “dengan” yang tidak memberikan informasi signifikan, sehingga analisis dapat lebih fokus pada kata-kata yang relevan. Selanjutnya, tahap stemming diterapkan untuk mengurangi variasi kata dengan mengubah kata-kata seperti “membaca,” “dibaca,” dan “membacakan” menjadi bentuk dasar “baca.” Tahap terakhir adalah tokenizing, di mana teks yang telah melalui proses sebelumnya dipecah menjadi unit kata terkecil sehingga dapat dikonversi ke dalam format yang sesuai untuk analisis sentimen.

Proses stemming dan tokenizing menjadi langkah penting dalam pengolahan data teks. Stemming bertujuan untuk menyederhanakan data dengan mengembalikan kata-kata seperti "membaca," "dibaca," dan "membacakan" ke bentuk dasarnya, yaitu "baca," sehingga analisis menjadi lebih efektif dan tidak rumit. Setelah itu, tahap tokenizing memecah teks yang telah diolah menjadi unit kata terkecil agar siap digunakan dalam analisis sentimen. Selanjutnya, data yang telah diproses melalui tahap transformasi dengan metode TF-IDF untuk menilai tingkat signifikansi suatu kata dalam dokumen tertentu dibandingkan dengan dokumen lainnya. TF-IDF memastikan bahwa analisis fokus pada kata-kata yang relevan untuk meningkatkan akurasi pengolahan.

Pada tahap data mining, algoritma Naive Bayes digunakan untuk mengklasifikasikan sentimen menjadi tiga kategori, yaitu positif, negatif, dan netral, dengan memanfaatkan fungsi MultinomialNB dari pustaka sklearn dan smoothing Laplace untuk mengatasi kemungkinan adanya nilai nol dalam data pelatihan. Setelah model dilatih dengan data pelatihan, prediksi dilakukan pada data uji untuk mendapatkan hasil klasifikasi. Proses ini diakhiri dengan evaluasi menggunakan matriks kebingungan untuk mengukur kinerja model, yang menghasilkan nilai akurasi, presisi, recall, skor F1, dan support. Dari hasil evaluasi, model berhasil mengidentifikasi 15.441 data sentimen positif, 847 data sentimen netral, dan 3.712 data sentimen negatif, memberikan wawasan mengenai skenario terbaik yang dihasilkan oleh algoritma.

```
[ ] 1 # Pembobotan TF-IDF
     2 tfidf_vectorizer = TfidfVectorizer()
     3 tfidf_train = tfidf_vectorizer.fit_transform(X_train)
     4 tfidf_test = tfidf_vectorizer.transform(X_test)
```

Gambar 2. Kode TF-IDF

```

1 # Tuning alpha untuk Naive Bayes
2 nb = MultinomialNB(alpha=0.25)
3 nb.fit(tfidf_train, y_train)
4 y_pred = nb.predict(tfidf_test)
    
```

Gambar 3. Kode Naive Bayes

```

Hasil Evaluasi Model:
Accuracy: 0.87315
Confusion Matrix:
[[ 2825   0  887]
 [  449   1  397]
 [  802   2 14637]]
    
```

	precision	recall	f1-score	support
Negatif	0.69	0.76	0.73	3712
Netral	0.33	0.00	0.00	847
Positif	0.92	0.95	0.93	15441
accuracy			0.87	20000
macro avg	0.65	0.57	0.55	20000
weighted avg	0.85	0.87	0.86	20000

Gambar 4. Hasil Evaluasi Model

SIMPULAN

Penelitian ini menganalisis sentimen ulasan aplikasi Shopee menggunakan algoritma Naive Bayes pada 100.000 data ulasan dari Google Play Store. Data diproses menjadi kategori sentimen positif, negatif, dan netral. Hasil evaluasi menunjukkan akurasi 87%, dengan kinerja terbaik pada sentimen positif (precision 92%, recall 95%, f1-score 93%) dan hasil yang memuaskan pada sentimen negatif (precision 69%, recall 76%, f1-score 73%). Meskipun demikian, algoritma Naive Bayes memiliki keterbatasan dalam mendeteksi sentimen netral karena jumlah data netral yang lebih kecil. Untuk hasil yang lebih optimal, disarankan untuk menambah data netral atau mempertimbangkan algoritma lain yang lebih efektif. Penelitian ini memberikan wawasan tentang pengalaman pengguna Shopee dan rekomendasi untuk penelitian selanjutnya dengan membandingkan algoritma lain, seperti SVM atau Decision Tree.

DAFTAR PUSTAKA

- [1] K. Cindy Pradhisa and R. Fajriyah, “Analisis Sentimen Ulasan Pengguna E-commerce di Google Play Store Menggunakan Metode IndoBERT,” *Technology and Science (BITS)*, vol. 6, no. 1, 2024, doi: 10.47065/bits.v6i1.5247.
- [2] D. P. Ray, F. N. Hasan, and A. R. Dzirkillah, “KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Terhadap KPU 2024 Berdasarkan Tweet Media Sosial Twitter Menggunakan Algoritma Naïve Bayes,” *Media Online*, vol. 4, no. 4, pp. 2235–2243, 2024, doi: 10.30865/klik.v4i4.1587.

- [3] D. S. Nugroho, I. F. Hanif, M. A. Hasbi, F. Fredianto, A. M. Saputra, and R. Zildjian, “Analisis Sentimen Dugaan Pelanggaran Pemilu 2024 Berdasarkan Tweet Menggunakan Algoritma Naïve Bayes Classifier,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 1169–1176, Jul. 2024, doi: 10.57152/malcom.v4i3.1496.
- [4] R. Kurniawan and Y. Arie Wijaya, “Analisis Data Sentimen Ulasan Pengguna Aplikasi Shopee di Google Play Store dengan Klasifikasi Algoritma Naïve Bayes,” vol. 6, no. 1, 2024.
- [5] A. Rizki Rinaldi, J. Perjuangan No, and B. Majasem Kec Kesambi Kota Cirebon, “PENERAPAN METODE NAÏVE BAYES CLASSIFIER PADA ANALISIS SENTIMEN APLIKASI GOPAY,” 2024.
- [6] A. Nurian, M. S. Ma’arif, I. N. Amalia, and C. Rozikin, “ANALISIS SENTIMEN PENGGUNA APLIKASI SHOPEE PADA SITUS GOOGLE PLAY MENGGUNAKAN NAIVE BAYES CLASSIFIER,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, Jan. 2024, doi: 10.23960/jitet.v12i1.3631.
- [7] R. A. Sitorus *et al.*, “Application of the Naïve Bayes Algorithm in Sentiment Analysis of Using the Shopee Application on the Play Store,” 2024, doi: 10.31849/digitalzone.v15i1.19828.