

ANALISA SENTIMEN ULASAN DI TOKOPEDIA DENGAN METODE NAIVE BAYES

Agus Gondowastardmojo^{1*}, Rajnaparamitha Kusumastuti²

¹Prodi S1 Informatika, STMIK Amikom Surakarta

²Sukoharjo Indonesia

Email: 1agus.10228@mhs.amikomsolo.ac.id, 2rajna@dosen.amikomsolo.ac.id

Abstract

This study analyzes user review sentiment on Tokopedia using the Naïve Bayes method. Data was sourced from the Kaggle platform and underwent several stages, including text normalization, tokenization, and stemming, to prepare it for further analysis. The Naïve Bayes model was implemented with TF-IDF and successfully classified user reviews into positive and negative sentiments with an accuracy rate of 90.54%. The analysis results show that this model effectively detects positive reviews with a precision of 90% and negative reviews with a precision of 100%. The study concludes that the Naïve Bayes method with TF-IDF can serve as a reliable tool for understanding user sentiment on e-commerce platforms like Tokopedia. It recommends further exploration with a more diverse dataset and the use of more advanced models to enhance sentiment analysis performance

Keywords: Sentiment Analysis, Naïve Bayes, TF-IDF, Tokopedia, User Reviews

Abstraksi

Penelitian ini menganalisis sentimen ulasan pengguna Tokopedia dengan menggunakan metode *Naïve Bayes*. Data yang digunakan diambil dari platform Kaggle dan melalui beberapa tahap, termasuk normalisasi teks, tokenisasi, dan stemming, untuk menyiapkan data untuk analisis lebih lanjut. Model *Naïve Bayes* diimplementasikan dengan teknik *TF-IDF* dan berhasil mengklasifikasikan ulasan pengguna menjadi sentimen positif dan negatif dengan tingkat akurasi sebesar 90,54%. Hasil analisis menunjukkan bahwa model ini efektif dalam mendeteksi ulasan positif dengan *precision* sebesar 90% dan ulasan negatif dengan *precision* mencapai 100%. Kesimpulan dari penelitian ini menunjukkan bahwa metode *Naïve Bayes* dengan *TF-IDF* dapat digunakan sebagai alat yang andal untuk memahami sentimen pengguna di platform *e-commerce* seperti Tokopedia. Penelitian ini merekomendasikan eksplorasi lebih lanjut dengan dataset yang lebih beragam dan penggunaan model yang lebih canggih untuk meningkatkan performa analisis sentimen

Kata Kunci: Analisis Sentimen, Naïve Bayes, TF-IDF, Tokopedia, Ulasan Pengguna

1. PENDAHULUAN

Tokopedia merupakan salah satu perusahaan e-commerce terbesar dan memiliki keunikan yang luar biasa dalam melakukan bisnisnya. *Features* yang terdapat di dalam tokopedia sangat membantu konsumen dalam berbelanja secara nyaman dan aman, dimana perkembangan tingkat keamanan transaksi terus diperbaharui dan dikembangkan untuk dapat memberikan yang terbaik kepada konsumen [1]. Tokopedia merupakan toko online yang paling banyak dikunjungi masyarakat Indonesia yakni dengan total pengunjung mencapai 1,2 miliar yang terbagi dari 863,1 juta pengunjung dari web mobile dan 329,8 juta pengunjung dari desktop [2]. Analisis sentimen adalah teknik pembersihan data dari kata dan simbol yang tidak relevan dan merubah data yang bernilai kualitatif menjadi data kuantitatif, kemudian data ulasan pengguna akan diklasifikasikan untuk mendapatkan ulasan yang bersifat positif dan negatif[3]. Algoritma *Naïve Bayes* adalah algoritma klasifikasi sederhana yang mana menghitung sekumpulan probabilitas dengan cara menjumlahkan dan mengkombinasikan nilai dari dataset yang diberikan[3].

Saat ini, situs-situs marketplace online menjadi pilihan favorit masyarakat untuk berbelanja produk karena menawarkan berbagai kelebihan, seperti kemudahan berbelanja tanpa harus datang langsung ke lokasi penjual, kebebasan wilayah, akses toko 24 jam, transaksi tanpa tawar-menawar, serta banyaknya diskon, promo, bahkan ongkos kirim gratis. Namun, kelebihan ini juga membawa beberapa kekurangan, seperti adanya risiko penipuan saat transaksi, barang yang dipesan tidak dikirim, perbedaan antara barang di foto dan yang diterima, kerusakan kemasan, hingga kesalahan warna produk. Oleh karena itu, konsumen sering memanfaatkan media sosial untuk menyampaikan keluhan mengenai pelayanan yang diberikan oleh perusahaan marketplace online tersebut[4]. Banyaknya interaksi ini menciptakan kebutuhan untuk memahami sentimen pengguna sebagai alat untuk meningkatkan kualitas layanan dan pengalaman pengguna. Namun, identifikasi dan klasifikasi sentimen ulasan secara manual menjadi tidak efisien mengingat jumlah data yang sangat besar[5]. Dalam aktivitas perdagangan online, sering kali ditemukan masalah terkait kualitas produk yang tidak memenuhi harapan pelanggan, yang pada akhirnya berdampak pada penurunan tingkat penjualan. Dalam konteks ini, platform e-commerce seperti Tokopedia menjadi salah satu media utama bagi pelanggan untuk menyampaikan pengalaman mereka melalui ulasan produk di IminOfficial Store. Oleh karena itu, analisis sentimen pelanggan berdasarkan ulasan di Tokopedia dapat memberikan wawasan berharga mengenai persepsi pelanggan terhadap kualitas produk yang ditawarkan di platform tersebut[6]. Masalah yang sering dihadapi oleh banyak perusahaan e-commerce di Indonesia saat ini adalah kesulitan dalam mendapatkan gambaran menyeluruh dari semua ulasan pelanggan. Hal ini disebabkan oleh waktu yang lama yang diperlukan untuk membaca setiap ulasan secara individu. Oleh karena itu, diperlukan suatu metode yang dapat merangkum opini pelanggan secara efektif, yaitu dengan menggunakan analisis sentimen. Menurut Kamus Besar Bahasa Indonesia (KBBI), sentimen diartikan sebagai pandangan atau pendapat yang didasari oleh perasaan yang

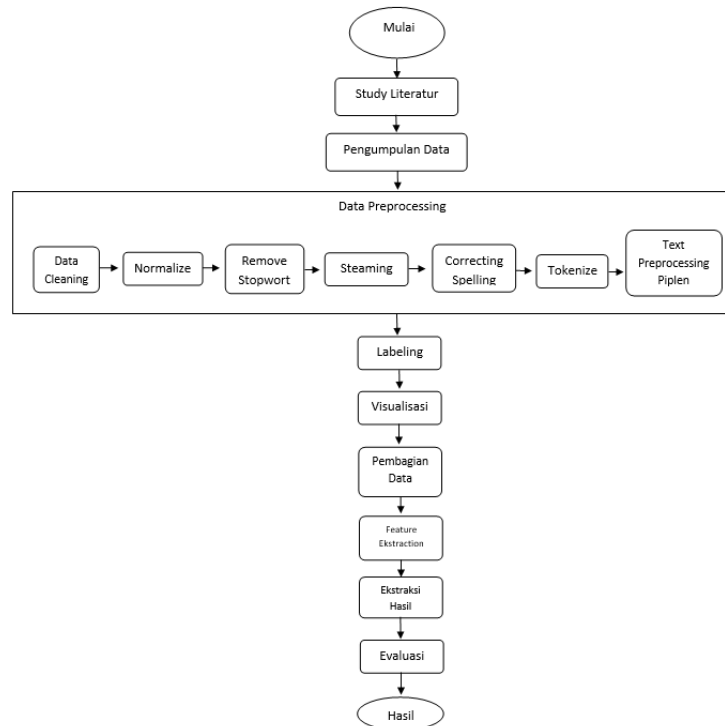
cenderung berlebihan terhadap sesuatu[7]. Oleh karena itu, diperlukan sebuah metode yang andal dalam analisis sentimen yang dapat mengolah data ulasan secara otomatis. Metode *Naïve Bayes* merupakan salah satu pendekatan yang banyak digunakan dalam analisis sentimen karena kemampuannya dalam mengklasifikasikan teks secara efisien. Algoritma *Naïve Bayes* menghitung probabilitas suatu teks menjadi positif atau negatif dengan menggunakan pendekatan statistik yang sederhana namun efektif[5]. Analisis sentimen dilakukan pada konten yang dibuat pengguna, seperti survei, media sosial, dan aplikasi terkait kesehatan. Layanan klien hingga perawatan klinis. Dengan pengembangan model RoBERTa, misalnya dapat menganalisis data domain yang lebih sulit dianalisis, seperti teks berita yang mungkin disertakan oleh penulis, pendapat/sentimen mereka kurang *eksplisit* [8]. Maka dari itu, eksplorasi lanjutan untuk metode yang lebih mutakhir, seperti penggunaan model pembelajaran mendalam (*deep learning*) atau pemrosesan bahasa alami yang lebih kompleks, perlu dilakukan untuk meningkatkan akurasi dan performa analisis sentimen dalam ulasan Tokopedia. Penelitian ini bertujuan untuk mengidentifikasi metode yang menghasilkan nilai terbaik untuk *Accuracy*, *Precision*, *Recall*, dan *Area Under the Curve (AUC)* di antara algoritma *Naïve Bayes*. [9]

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan berbagai algoritma untuk analisa sentimen di Tokopedia. *Algoritma Multinomial Naïve Bayes* digunakan untuk mengklasifikasikan opini pelanggan menjadi kategori positif dan negatif, sehingga dapat mengidentifikasi faktor-faktor yang memengaruhi kepuasan pelanggan di *e-commerce* Tokopedia[7]. Studi menyediakan ini solusi untuk menentukan proporsi tanggapan pengguna aplikasi terhadap Tokopedia, yang dapat digunakan untuk meningkatkan kualitas layanan, kualitas proporsi, dan operasi internal perusahaan. Tanggapan pengguna aplikasi terhadap Tokopedia, yang dapat digunakan untuk meningkatkan kualitas layanan, kualitas produk, dan operasi internal perusahaan [10]. Tujuan studi ini untuk membuat aplikasi yang mampu mengkategorikan yang mampu data dari berbagai sumber menjadi tiga kategori: positif, negatif, dan netral. Direkomendasikan untuk digunakan metode *Naïve Bayes* untuk mengembangkan aplikasi ini karena menurut beberapa literatur, dapat menghitung ambang batas kualitas tinggi[11]. Hasil analisis sentimen ini nantinya dapat digunakan sebagai bahan evaluasi untuk meningkatkan kualitas produk. Dengan produk yang berkualitas baik, ulasan dari pembeli akan menjadi lebih positif. Hal ini dapat meningkatkan citra perusahaan di mata masyarakat, sehingga minat beli konsumen dan pendapatan perusahaan pun ikut meningkat[12].

2. METODE PENELITIAN

2.1. Alur Penelitian

Bagian ini berisi penjelasan prosedur dan urutan langkah-langkah penelitian yang divisualisasikan dengan gambar 1 bagian alur penelitian (*flowchart*). Setiap tahapan penelitian pada alur disertai dengan uraian penjelasan.



Gambar 1. Alur Penelitian

2.2. Study Literatur

Studi literatur ini mendalami analisis sentimen sebagai bagian dari pemrosesan bahasa alami (*NLP*) untuk mengenali sikap atau opini dalam teks. Proses ini mencakup pengumpulan data, pra-pemrosesan, representasi teks, pelabelan, pelatihan model, evaluasi, dan prediksi sentimen ulasan Tokopedia dengan teks singkat dan bahasa informal, merupakan *platform* populer untuk analisis sentimen. *Naive Bayes*, algoritma sederhana namun efisien berbasis *Teorema Bayes*, cocok untuk klasifikasi teks dengan varian seperti *Multinomial*, *Bernoulli*, dan *Gaussian*. Studi ini menyediakan dasar teori dan metode untuk membangun model yang efektif dalam mengklasifikasikan sentimen ulasan pada Tokopedia.

2.3. Pengumpulan Data

Data dalam penelitian ini dataset diambil dari *Kaggle*. Pengolahan dan analisis data dilakukan menggunakan bahasa pemrograman *Python*. Hasil analisis ini bertujuan untuk menentukan ulasan Positif dan Negatif. Dataset ini diambil di *Kaggle* dengan data sebanyak 300 dengan 5 kolom yang berisi : Nama_Produk, Akun, Ulasan_Clear, Label. [13]

2.4. Preprocessing

Preprocessing adalah salah satu teknik yang biasanya digunakan dalam basis data adalah persiapan data, yang melibatkan penghapusan beberapa data yang tidak konsisten, konsisten, missing value, dan noise[14]. Misalnya, "yg" diubah menjadi "yang" dan "thx" menjadi "terima kasih". Proses ini diterapkan pada kolom 'Ulasan_Clean' dalam *DataFrame* `df`, sehingga teks lebih konsisten dan siap untuk analisis. Selain itu, kata-kata pengisi dihilangkan dari kolom tersebut untuk membersihkan teks. Selanjutnya, teks dalam kolom 'Ulasan_Clean' dipecah menjadi daftar kata melalui proses tokenisasi, yang penting untuk analisis lebih mendalam. Proses *stemming* juga dilakukan untuk mengubah kata-kata ke bentuk dasarnya, dengan hasil akhir disimpan dalam file CSV di Google Drive. Dengan langkah-langkah ini, *DataFrame* `df` sekarang berisi teks yang sudah dinormalisasi, dibersihkan, dan di-*stemming*, sehingga siap untuk analisis lebih lanjut seperti pencarian kata kunci, klasifikasi teks, atau analisis sentimen.

2.5. Labeling

Perintah `!pip install` digunakan untuk menginstal pustaka seperti `textblob`, `sastrawi` dan `wordcloud`, `sastrawi` untuk stemmer atau pengubah kata berimbuhan (kata turunan) ke bentuk dasar, `textBlob` digunakan untuk menganalisis sentimen ulasan ke dalam kategori positif atau negatif, `wordcloud` untuk membantu merangkum informasi teks yang besar menjadi bentuk visual yang mudah dipahami

2.6. Visualisasi

Langkah pertama melibatkan impor kelas `WordCloud` untuk membuat *word cloud* dan `STOPWORDS` untuk menghilangkan kata-kata umum. Peneliti kemudian mendefinisikan fungsi `plot_cloud(wordcloud)` untuk menampilkan visualisasi. Seluruh teks dalam kolom 'Ulasan_Clean' dari *DataFrame* `my_df` digabungkan menjadi satu string dan disimpan dalam variabel `all_words`. Lalu, objek `WordCloud` dibuat dengan parameter tertentu, seperti ukuran dan warna latar belakang, serta pengaturan agar tidak menampilkan kolokasi kata. Kata-kata umum dihapus menggunakan `STOPWORDS`. Dengan memanggil `generate(all_words)`, *word cloud* dibentuk berdasarkan frekuensi kata, dan fungsi `plot_cloud(wordcloud)` menampilkan visualisasi tersebut, memudahkan pemahaman pola dalam data.

2.7. Pembagian Data

Kode ini mempersiapkan data untuk analisis sentimen dengan machine learning. Pertama, kolom ulasan bersih (`Ulasan_clean`) disimpan di `X` sebagai fitur, sementara kolom `sentimen` (positif atau negatif) disimpan di `y` sebagai target. `LabelEncoder` mengubah label sentimen menjadi angka agar dapat digunakan dalam model. Data kemudian dibagi menjadi data latih (70%) dan data uji (30%) menggunakan `train_test_split`, dengan `random_state=42` untuk memastikan hasil pembagian data yang konsisten. Langkah-langkah ini menyiapkan data untuk pelatihan model klasifikasi.

2.8. Fiture Extraction

Kode ini menggunakan metode *TF-IDF* (*Term Frequency-Inverse Document Frequency*) untuk mengubah teks ulasan menjadi fitur numerik, yang memudahkan pemrosesan oleh model *machine learning*. Pertama, objek `TfidfVectorizer` dari `sklearn.feature_extraction.text` diinisialisasi dan digunakan untuk mengonversi teks ulasan (`X`) ke dalam matriks *TF-IDF* dengan fungsi `fit_transform(X)`. Proses ini menghitung frekuensi setiap kata dalam teks serta membandingkannya dengan kemunculannya di dokumen lain, sehingga memberikan bobot yang lebih tinggi pada kata-kata yang spesifik untuk dokumen tertentu dan lebih rendah untuk kata-kata umum. Hasilnya, variabel `X_tfidf` menyimpan representasi numerik dari teks ulasan dalam bentuk matriks yang siap digunakan sebagai input dalam model klasifikasi. Matriks ini ditampilkan dengan fungsi `print(X_tfidf)` untuk melihat hasil transformasi *TF-IDF* dari teks yang telah diproses.

2.9. Extraksi Hasil

Kode ini membangun model klasifikasi *Naive Bayes* menggunakan *Multinomial Naive Bayes* yang diterapkan pada data yang telah diproses dengan metode *TF-IDF*. Pertama, objek (`MultinomialNB`) dari (`sklearn.naive_bayes`) diinisialisasi untuk membuat model klasifikasi yang sesuai untuk data yang direpresentasikan sebagai matriks *TF-IDF*, yang bekerja baik untuk data berbentuk frekuensi atau probabilitas seperti hasil dari (`TfidfVectorizer`). Selanjutnya, model ini dilatih pada data ulasan yang telah dikonversi ke bentuk *TF-IDF* (`X_tfidf`) dan label sentimen (`y`) dengan metode `fit(X_tfidf, y)`. Proses pelatihan ini memungkinkan model untuk mempelajari pola dalam data yang sesuai dengan klasifikasi positif atau negatif, sehingga model dapat melakukan prediksi pada data ulasan baru dengan tingkat akurasi yang diharapkan tinggi.

2.10. Evaluasi

Kode ini digunakan untuk mengukur kinerja model *Naive Bayes* yang telah dilatih menggunakan fitur *TF-IDF* dalam klasifikasi sentimen. Pertama, model menghasilkan prediksi kelas sentimen pada data ulasan melalui fungsi (`predict`), dengan hasil prediksi disimpan dalam variabel (`y_pred_tfidf`). Prediksi ini dibandingkan dengan label sesungguhnya (`y`) menggunakan `accuracy_score`, yang menghitung proporsi prediksi benar untuk mendapatkan nilai akurasi, disimpan dalam variabel (`accuracy_tfidf`). Selanjutnya, kode `print(classification_report(y, y_pred_tfidf))` menampilkan laporan klasifikasi yang merinci kinerja model berdasarkan metrik *Precision*, *Recall*, dan *F1-Score* untuk setiap kelas (positif dan negatif). Laporan ini memberikan gambaran menyeluruh mengenai efektivitas model dalam mengklasifikasikan sentimen, sehingga evaluasi kinerja model tidak hanya bergantung pada akurasi tetapi juga pada kemampuan mendeteksi kelas secara spesifik.

3. HASIL DAN PEMBAHASAN

Pada gambar 2 akan menyajikan hasil dari penelitian yang telah dilakukan. Data penelitian diambil dari website *Kaggle*, dengan total data terkumpul 300 data yang terdiri 4 data yaitu : Nama_Produk, Akun, Ulasan_Clean, Label, untuk dibagian label terdiri 2 angka yaitu 0 untuk ngeatif 1 untuk positif. Maka terbentuklah dataset seperti dibawah ini :

A	B	C	D	E	F
	Nama_Produk	Akun	Ulasan_clean	label	
1					
2	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Stephanus	dua kali beli sayang lambat respon order pagi kirim sore respon respon itu kirim besok barang bagus sesuai harga	0	
3	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Yati	bagus kirim cepat	1	
4	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Husin	barang sesuai pesan	1	
5	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Febby	panas pakai	1	
6	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Rahmat	barang sesuai minta bagus komunikasi jual beli pro aktif	0	
7	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Alden	bagus masker	1	
8	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Heni	pengiriman cepat produk bagus konsumen suka makasih	1	
9	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Husin	barang sesuai pesan kirim cepat	1	
10	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	n***i	proses barang jemput kurir banget anjias pokok	0	
11	Masker KF94 4 Ply Isi 10 Pcs Masker Medis Kesehatan Kemasan Plastik - Putih	Antonius	mantabspesan warna hitam putih pink	1	
12	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - M***e		kirim cepat sesuai pesan respon jual cepat ramah	1	
13	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - R***i		barang terima sesuai pesan aman	1	
14	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Frida		tidak ada komentar	1	
15	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Renold		kecewa berat masker hijab warna hitam tali masker gampang putus tali masker pendek	0	
16	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Roy		kirim sesuai estimasi baik	1	
17	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Ridwan		karet headloop nya pendek istri pakai untung anak	0	
18	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Fatma		barang bagus hrga murah layan muas	1	
19	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - v***a		bagus barang sesuai	1	
20	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - A***s		tidak ada komentar	1	
21	MASKER 3PLY IZIN KEMENKES MASKER MEDIS KESEHATAN 3 PLY ALKINDO 50 PCS - Renold		masker medis lumayan cepat tuju harga murah barang murahansmoga masker medis nya manfaat	1	
22	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear Parama		sesuai pesan packing kirim cepat terima kasih ya	1	
23	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear S***i		sesuai harga	1	
24	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear Galih		kirim cepat barang sesuai	1	
25	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear Sujatmiko		sesuai pesan sip	1	
26	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear A***s		nice one	1	
27	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear Hae		sesuai pesan	1	
28	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear M***i		barang rumah	1	
29	Masker 3 ply Medis Masker Bedah Kemenkes 3ply isi 50pcs Merk Orlee - Masker Ear Achmad		proses nya ceoat thanks aan	1	

Gambar 2. Sample dataset yang digunakan

3.1. Clening Dataset

Pada gambar 3 adalah hasil dari Clening Dataset

	Ulasan_clean	label
0	dua kali beli sayang lambat respon order pagi ...	0
1	bagu kirim cepat	1
2	barang sesuai pesan	1
3	pana pakai	1
4	barang sesuai minta bagu komunikasi jual beli ...	0
...
295	barang kirim sesuai deskripsi tera ijin kemenk...	0
296	kotak penyok sana mngkn expediai	1
297	barang sesuai kualitas kardus penyok next time o...	1
298	harga promo murah kirim promo	1
299	prose kirim sabar ya temanteman beli paket cum...	1

300 rows × 2 columns

Gambar 3. Hasil cleaning dataset

Pada gambar 3 merupakan hasil pembersihan pada proses tersebut, *DataFrame* (*df*) telah dimodifikasi sehingga teks dalam kolom 'Ulasan_Clean' menjadi lebih bersih dan seragam, siap untuk digunakan dalam analisis lebih lanjut seperti pemrosesan teks atau analisis sentimen.

3.2. Tokenize

Gambar 4 dibawah ini menampilkan hasil Tokenize yang peneliti buat.

	Ulasan_clean
0	[kali, beli, sayang, lambat, respon, order, pa...
1	[bagu, kirim, cepat]
2	[barang, sesuai, pesan]
3	[pana, pakai]
4	[barang, sesuai, minta, bagu, komunikasi, jual...
...	...
295	[barang, kirim, sesuai, deskripsi, tera, ijin,...
296	[kotak, penyok, sana, mngkn, expediai]
297	[barang, sesuai, kualita, kardu, penyok, next,...
298	[harga, promo, murah, kirim, promo]
299	[prose, kirim, sabar, temanteman, beli, paket,...

300 rows x 1 columns

Gambar 4. Tokenize

Pada gambar 4 Tabel di atas menampilkan hasil *preprocessing* data teks pada kolom yang dinamakan "Ulasan_clean," yang berisi ulasan yang telah dibersihkan. Data yang disajikan merupakan hasil dari proses pembersihan teks, di mana setiap ulasan diubah menjadi kumpulan kata kunci (token) yang relevan dengan menghilangkan kata-kata yang tidak diperlukan, seperti *stop words* atau kata-kata yang tidak memiliki nilai informasi. Selain itu, mungkin dilakukan teknik-teknik lain seperti *stemming* atau *lemmatization* untuk menyederhanakan setiap kata ke bentuk dasarnya. Setiap baris dalam tabel berisi daftar kata yang merupakan hasil *tokenisasi* dari sebuah ulasan, di mana kata-kata tersebut telah difilter untuk menghilangkan karakter atau kata yang tidak bermakna dalam analisis. Contohnya, pada indeks ke-0, terdapat kata-kata seperti "kali," "beli," "sayang," dan "lambat" yang dapat memberikan indikasi tertentu mengenai sentimen atau maksud pengguna dalam ulasannya. Tabel ini terdiri dari 300 baris ulasan, masing-masing mewakili satu entri ulasan yang telah diproses untuk analisis selanjutnya, seperti analisis sentimen atau klasifikasi teks, di mana teks bersih ini memudahkan proses ekstraksi fitur menggunakan metode seperti *TF-IDF*.

3.3. Labeling

Gambar 5 dibawah ini adalah hasil dari Labeling yang telah dibuat.

	Ulasan_clean	label	sentimen
0	kali beli sayang lambat respon order pagi kiri...	0	negatif
1	bagu kirim cepat	1	positif
2	barang sesuai pesan	1	positif
3	pana pakai	1	positif
4	barang sesuai minta bagu komunikasi jual beli ...	0	negatif
...
295	barang kirim sesuai deskripsi tera ijin kemenk...	0	negatif
296	kotak penyok sana mngkn expediai	1	positif
297	barang sesuai kualita kardu penyok next time o...	1	positif
298	harga promo murah kirim promo	1	positif
299	prose kirim sabar temanteman beli paket cuman ...	1	positif

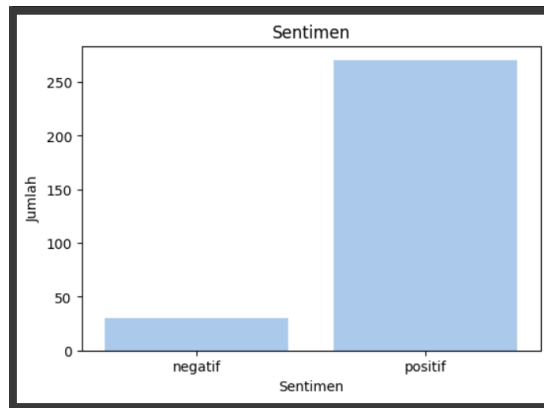
300 rows × 3 columns

Gambar 5. Labeling

Pada Gambar 5 ini menampilkan contoh data yang terdiri dari tiga kolom: Ulasan_clean, label, dan sentimen. Kolom Ulasan_clean berisi teks ulasan pelanggan yang telah dibersihkan, yang mencerminkan opini atau pengalaman mereka terhadap produk atau layanan tertentu. Kolom label menunjukkan nilai numerik, yaitu 0 dan 1, sebagai representasi kategori sentimen; di mana 0 mewakili sentimen negatif, dan 1 mewakili sentimen positif. Sementara itu, kolom sentimen berisi label kategorikal dalam bentuk teks ("negatif" atau "positif") untuk menggambarkan jenis sentimen dari ulasan tersebut. Data ini memiliki 300 baris, di mana setiap baris mewakili satu ulasan pelanggan. Struktur data ini memungkinkan untuk dilakukan analisis lebih lanjut, seperti klasifikasi sentimen menggunakan model machine learning. Sebelum melakukan analisis tersebut, kolom label dan sentimen dapat digunakan untuk melatih model dan mengevaluasi performa model dalam memprediksi sentimen ulasan pelanggan.

3.4. Visualisasi Labeling Data

Pada gambar 6 dibawah ini adalah hasil dari Visualisasi Labeling Data



Gambar 6. Hasil Visualisasi Labeling Data

Pada Gambar 6 Grafik batang di atas menunjukkan distribusi jumlah sentimen positif dan negatif dari data yang telah diklasifikasikan. Dalam grafik ini, sumbu horizontal (sumbu x) mewakili jenis sentimen, yaitu "positif" dan "negatif," sedangkan sumbu vertikal (sumbu y) menunjukkan jumlah data untuk masing-masing kategori sentimen. Warna biru muda pada batang memberikan visualisasi yang menunjukkan frekuensi lebih tinggi pada sentimen positif dibandingkan sentimen negatif. Hal ini mungkin menunjukkan bahwa algoritma klasifikasi yang digunakan mendeteksi lebih banyak sentimen positif dalam dataset yang diolah. Secara umum, grafik ini berfungsi untuk memberikan gambaran awal mengenai persebaran sentimen yang dihasilkan dari analisis sentimen, yang dapat menjadi indikator penting dalam memahami dinamika opini publik di media sosial, terutama dalam konteks tertentu seperti ulasan Tokopedia. Grafik menunjukkan ketidakseimbangan antara data sentimen positif dan negatif, dengan sentimen positif jauh lebih dominan. Ketidakseimbangan ini mungkin disebabkan oleh distribusi data yang tidak merata, bias dalam pengumpulan data, atau topik yang umumnya mendapat respons positif. Misalnya, data dari ulasan produk populer atau diskusi media sosial cenderung memiliki lebih banyak sentimen positif. Dalam analisis sentimen, ketidakseimbangan ini perlu diperhatikan, karena dapat mempengaruhi hasil, sehingga diperlukan metode khusus untuk menjaga akurasi dan representativitas analisis.

3.5. Visualisasi

Pada gambar 7 dibawah ini adalah hasil dari Visualisasi



Gambar 7. Hasil visualisasi

Pada gambar 7 menjelaskan *Word cloud* yang ditampilkan, dibuat menggunakan pustaka *Python WordCloud*, memvisualisasikan frekuensi kata dalam kolom 'full_text' dari dataset. Kata-kata yang lebih sering muncul ditampilkan dalam ukuran yang lebih besar, sementara yang kurang sering muncul tampil lebih kecil. Dengan latar belakang putih dan warna biru dan kuning, *word cloud* menyoroti kata-kata utama seperti sesuai, cepat, barang, bagus dan kirim. Ini memberikan wawasan tentang topik utama yang dibahas, serta membantu dalam analisis teks awal dengan menunjukkan kata-kata yang dominan dan pola-pola utama dengan cara yang cepat dan efisien.

3.6. Evaluasi

3.6.1. Classification dan TF-IDF

Pada gambar 8 adalah hasil evaluasi Naïve Bayes model TF-IDF

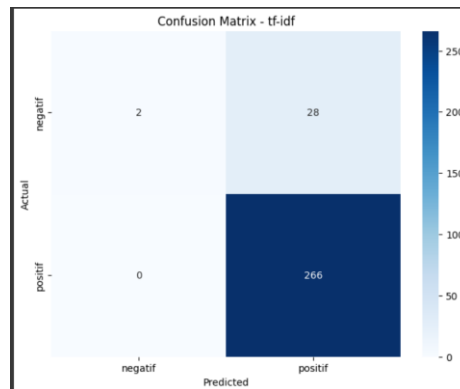
Akurasi model Naive Bayes dengan TF-IDF: 0.9054054054054054					
	precision	recall	f1-score	support	
0	1.00	0.07	0.12	30	
1	0.90	1.00	0.95	266	
accuracy			0.91	296	
macro avg	0.95	0.53	0.54	296	
weighted avg	0.91	0.91	0.87	296	

Gambar 8. Hasil evaluasi Naïve Bayes TF-IDF

Pada gambar 6 adalah hasil evaluasi *Naïve Bayes* model *TF-IDF*, dengan tingkat akurasi sebesar 90.54% (0.9054), kategori negatif (kelas 0) memiliki *precision* 100% dan *recall* 7%, yang menunjukkan kemampuan model untuk mendeteksi dan memprediksi ulasan negatif secara akurat. Sementara itu, pada kategori positif (kelas 1), *precision* mencapai 90% *recall* 100%, untuk rata – rata (*macro dan weighted*) *Macro Avg Precision* 0.95, *Recall* 0.53, dan *F1-score* 0.54. Rata-rata makro menunjukkan kinerja rata-rata di kedua kelas, tanpa memperhatikan proporsi masing-masing kelas, sedangkan *Weighted Avg Precision* 0.91, dan *Recall* 0.91, dan *F1-score* 0.87. Rata-rata berbobot menunjukkan kinerja keseluruhan model, dengan mempertimbangkan proporsi kelas.

3.6.2. Confusion matrik Model TF-IDF

Pada gambar 9 adalah hasil dari Confusion matrik Model TF-IDF



Gambar 9. Confusion Matrik Model TF-IDF

Pada gambar 9 Gambar di atas menampilkan confusion matrix dari hasil klasifikasi sentimen menggunakan metode *TF-IDF* sebagai representasi teks. Matriks ini menunjukkan jumlah prediksi yang benar dan salah dari model terhadap dua kelas sentimen, yaitu "positif" dan "negatif." Pada sumbu vertikal (*Actual*) tertera label asli dari data, sementara sumbu horizontal (*Predicted*) menunjukkan label prediksi dari model. Dari matriks ini, dapat dilihat bahwa model berhasil memprediksi 266 ulasan positif dengan benar, tetapi terdapat 28 ulasan negatif yang salah diklasifikasikan sebagai positif. Sementara itu, hanya ada 2 ulasan negatif yang diklasifikasikan dengan benar, dan tidak ada ulasan positif yang salah diklasifikasikan sebagai negatif. *Confusion matrix* ini memperlihatkan bahwa model cenderung sangat baik dalam mendeteksi sentimen positif, tetapi kurang efektif dalam mengenali ulasan dengan sentimen negatif. Ketidakseimbangan ini mengindikasikan bahwa representasi *TF-IDF* mungkin kurang optimal dalam menangkap fitur yang relevan untuk mengidentifikasi ulasan negatif, atau bahwa model cenderung bias terhadap kelas positif.

4. KESIMPULAN

Penelitian ini menganalisis sentimen ulasan pengguna Tokopedia menggunakan metode *Naïve Bayes*. Hasilnya menunjukkan bahwa algoritma *Naïve Bayes* dengan teknik TF-IDF berhasil mengklasifikasikan ulasan sebagai positif atau negatif dengan akurasi sebesar 90,54%. Kategori positif memiliki *precision* tinggi, yaitu 90%, sedangkan kategori negatif mencapai *precision* 100%. *Confusion matrix* menunjukkan 266 ulasan positif terklasifikasi dengan benar dan hanya terdapat beberapa kesalahan klasifikasi. Secara keseluruhan, metode ini terbukti efektif untuk mengidentifikasi sentimen pengguna pada platform e-commerce seperti Tokopedia. Rekomendasi untuk penelitian selanjutnya mencakup pengembangan model yang lebih canggih untuk meningkatkan akurasi klasifikasi serta eksplorasi metode lain untuk optimasi analisis sentimen.

5. SARAN

Penelitian ini masih memiliki beberapa kekurangan seperti penggunaan data yang lebih beragam, penelitian selanjutnya dapat memanfaatkan dataset yang lebih besar dan beragam dari sumber lain selain *Kaggle*, seperti langsung dari ulasan aplikasi e-commerce lain atau media sosial, agar hasil analisis lebih representatif. Pendekatan multikategori sentimen, selain klasifikasi positif dan negatif, sebaiknya dikembangkan kategori sentimen tambahan seperti netral atau sangat positif atau negatif untuk menangkap spektrum opini yang lebih luas dari pengguna. Untuk mengatasi Ketidakseimbangan data dalam analisis sentimen, seperti dominasi kelas positif pada grafik, dapat menyebabkan model bias terhadap kelas mayoritas, sehingga mengabaikan kelas minoritas. Untuk mengatasi hal ini, beberapa metode dapat diterapkan. *Oversampling* pada kelas minoritas, misalnya dengan *SMOTE (Synthetic Minority Over-sampling Technique)*, dapat meningkatkan jumlah data pada kelas negatif. Alternatifnya, *undersampling* pada kelas mayoritas dapat membantu menyeimbangkan distribusi data, meski berisiko kehilangan informasi. Pendekatan lain termasuk penentuan bobot kelas agar model lebih fokus pada kelas minoritas dan penggunaan *ensemble methods* seperti *Balanced Random Forest*. Algoritma khusus seperti *XGBoost* juga menyediakan parameter untuk menangani ketidakseimbangan ini. Pemilihan teknik yang tepat dan validasi silang diperlukan agar model tetap akurat dan representatif.

DAFTAR PUSTAKA

- [1] I. Gamayanto, "Inovasi & Pengembangan E-Commerce - Crm Dalam Menghadapi Globalisasi: Studi Kasus -Tokopedia.Com," *J. Sist. Inf. Dan Teknol. Inf.*, vol. 7, no. 1, pp. 89–103, 2018.
- [2] R. Apriani and D. Gustian, "Analisis Sentimen Dengan Naïve Bayes Terhadap Komentar Aplikasi Tokopedia," *J. Rekayasa Teknol. Nusa Putra*, vol. 6, no. 1, pp. 54–62, 2019, doi: 10.52005/rekayasa.v6i1.86.
- [3] Rahel Lina Simanjuntak, Theresia Romauli Siagian, Vina Anggriani, and Arnita Arnita, "Analisis Sentimen Ulasan Pada Aplikasi E-Commerce Shopee Dengan Menggunakan Algoritma Naïve Bayes," *J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 3, no. 3, pp. 23–39, 2023, doi: 10.55606/teknik.v3i3.2411.
- [4] D. P. Artanti, A. Syukur, A. Prihandono, and D. R. I. M. Setiadi, "Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," pp. 8–9, 2018.
- [5] L. Dan and B. Menggunakan, "Analisis sentimen aplikasi shopee, tokopedia, lazada

- dan blibli menggunakan leksikon dan random forest,” vol. 12, no. 3, 2024.
- [6] R. T. S. Syahjuddin Azra, “Analisis Sentimen Menggunakan Metode Naïve Bayes Terhadap Produk Pt . Imin Technology Berdasarkan Sentiment Analysis Using the Naïve Bayes Method on Pt . Imin Technology Products Based on Reviews From,” vol. 2, no. September, pp. 445–452, 2023.
- [7] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, “Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelanggan Tokopedia,” *J. Gaussian*, vol. 11, no. 4, pp. 562–571, 2023, doi: 10.14710/j.gauss.11.4.562-571.
- [8] D. Hermawan, M. Akhsanal, Z. Wahyudi, A. Ariyanto, and D. Dwi, “05. Optimasi Analisis Sentimen Pada Twitter Olshop,” vol. 6, no. September, pp. 821–828, 2022, [Online]. Available: <http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/493>
- [9] ANANDA MUHAMAD TRI UTAMA, “Komparasi Algoritma Klasifikasi Text Mining Untuk Analisa Sentimen Pada Akun Twitter Tokopediacare,” vol. 9, no. 1, pp. 356–363, 2022.
- [10] S. M. Salsabila, A. Alim Murtopo, and N. Fadhilah, “Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier,” *J. Minfo Polgan*, vol. 11, no. 2, pp. 30–35, 2022, doi: 10.33395/jmp.v11i2.11640.
- [11] S. Hanggara, T. M. Akhriza, and M. Husni, “Aplikasi Web Untuk Analisis Sentimen Pada Opini Produk Dengan Metode Naive Bayes Classifier,” *Semin. Nas. Inov. Dan Apl. Teknol. Di Ind.*, vol. 4, no. 2, p. A33.1-A33.6, 2017.
- [12] N. Hasyimiyyah, I. A. Mashudi, R. Arianto, J. T. Informasi, and P. N. Malang, “NAÏVE BAYES CLASSIFIER,” pp. 9–18, 2024.
- [13] Raniapriliyap, “No Title,” kaggle. [Online]. Available: <https://www.kaggle.com/code/raniaprillyap/analisa-sentimen-ulasan-di-tokopedia-2/notebook>
- [14] L. C. Defitri, “Normalisasi Data Siswa Dengan Model Desimal Scaling Untuk Siswa Menerima Bantuan Dengan Algoritma K-Medoids,” *Journal.Stmik-Muliadarma.Ac.Id*, vol. 1, no. 2, pp. 55–59, 2024.