Analisis Ulasan Produk Menggunakan Metode *Naive Bayes Classifier*

Monika Hermiani Yolanda Simamora[a1], Ida Bagus Made Mahendra[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Udayana, Bali
Jln. Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, 08261, Bali, Indonesia
[1]monikasimamora8@gmail.com
[2]ibm.mahendra@unud.ac.id

Abstract

Advancements in technology have shifted market activities towards e-commerce, resulting in a substantial increase in user-generated review data. Buyer reviews, which are comments provided after purchasing products online, serve as valuable feedback for sellers to enhance product quality and aid buyers in making informed decisions. However, manually analyzing a large volume of buyer reviews is time-consuming. To address this issue, sentiment analysis methods can be employed to automatically classify product reviews into positive and negative sentiment classes.. Sentiment analysis was conducted using Multinomial Naive Bayes in this study.. The data used were 400 pieces of data with a division of 80% as training data and 20% as test data. The preprocessing in this study are data cleaning, tokenization, normalization, stopword, and stemming. The feature extraction process is carried out by the Term-Frequency method. . Then the classification process is carried out using the Multinomial Naive Bayes method and tested using the Confusion Matrix method. The final results of this study showed that the Multinomial Naive Bayes method could carry out the product review data classification process well and obtained an accuracy value of 85%, a precision value of 77%, a recall value of 72%, and an f1-score value of 74%.

Keywords: *Text preprocessing, Term Frequency, Naive Bayes Classifier, Confusion Matrix*

1. Pendahuluan

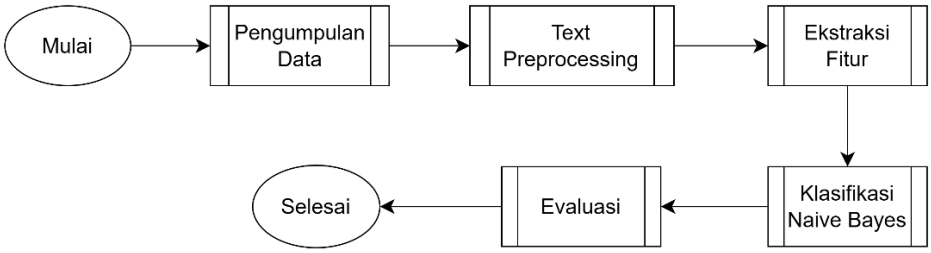
Teknologi canggih yang berkembang pada masa kini menyebabkan kehidupan manusia dalam berbagai aspek mulai berubah. Salah satu bidang yang terasa sangat jelas perubahannya adalah bidang ekonomi. Masyarakat yang awalnya melakukan kegiatan jual beli secara konvensional di pasar mulai beralih ke *e-commerce* yang dapat dilakukan secara *online*. Hal ini dibuktikan dengan jumlah pengguna *e-commerce Shopee* di Indonesia mencapai 158 juta pada kuartal I 2023 [1]. Besarnya jumlah pengguna *e-commerce* tersebut tentu akan menghasilkan data ulasan pembeli dalam jumlah yang besar pula. Ulasan pembeli dapat berupa komentar atau *feedback* yang diberikan setelah melakukan pembelian pada *e-commerce.* Ulasan pembeli tersebut dapat dianalisis dan hasil analisis tersebut dapat digunakan sebagai bahan evaluasi untuk peningkatan kualitas produk maupun pelayanan dari penjual serta membantu pembeli melakukan penilaian sebelum membeli produk. Menganalisis secara manual data ulasan pembeli yang berjumlah besar tentu akan menghabiskan banyak waktu dan tidak efisien sehingga dibutuhkan sebuah sistem yang dapat melakukan analisis tersebut secara otomatis. Sistem tersebut disebut dengan analisis sentimen.

Analisis sentimen merupakan sebuah proses mengekstraksi dan mengidentifikasi sentimen atau opini berupa data teks dan mengklasifikasikannya menjadi kelas sentimen positif atau kelas sentimen negatif. Proses pengklasifikasian sentimen dilakukan menggunakan metode klasifikasi. Salah satu metode klasifikasi yang banyak digunakan adalah metode *Naive Bayes Classifier.* Algoritma *Naive Bayes Classifier* (NBC) merupakan metode klasifikasi yang didasarkan pada perhitungan probabilitas sesuai dengan teorema *Bayes* [3].

Pada penelitian Febry [4] membuat sistem analisis sentimen untuk mengklasifikasikan sentimen SARA, hoaks, dan radikal pada postingan media sosial. Penelitian tersebut berhasil mengklasifikasikan sentimen secara akurat dengan nilai akurasi sebesar 99,62%. Penelitian yang dilakukan adalah analisis terhadap ulasan atau sentimen produk pada salah satu *e-commerce* yaitu *Shopee* dengan menggunakan metode *Naive Bayes Classifier* (NBC). Penelitian ini bertujuan untuk mengklasifikasikan ulasan pembeli menjadi kelas sentimen positif dan negatif.

2. Metode Penelitian

Proses pada penelitian ini dilakukan melalui beberapa tahapan proses antara lain proses pengumpulan data, *preprocessing*, ekstraksi fitur, klasifikasi menggunakan metode *Naive Bayes,* dan yang terakhir adalah pengujian. Berikut ini adalah flowchart dari metode penelitian yang akan dilakukan :



Gambar 1. Flowchart Metode Penelitian

2.1 Pengumpulan Data

Penelitian ini menggunakan data primer berupa ulasan produk Lampu 3D LED dari *Shopee*. Ulasan diambil secara acak menggunakan metode *web scraping* yang dilakukan dengan bahasa pemrograman Python. Data yang digunakan akan diberi label positif dan negatif berdasarkan rating pada ulasan produk.

2.2 Text Preprocessing

Preprocessing merupakan proses mengolah data awal pada teks untuk mempersiapkan teks menjadi data yang dapat diolah [2]. Beberapa tahapan yang digunakan pada preprocessing ini adalah sebagai berikut:

a. *Cleaning Data*
   Pada proses *cleaning* dilakukan penghapusan karakter seperti tanda baca, simbol, angka,dan lainnya. Pada proses ini juga dilakukan proses penyeragaman semua huruf menjadi huruf kecil.

b. *Tokenization*
   Proses ini memotong kalimat pada kata menjadi satuan kata.

c. *Normalization*
   Proses ini dilakukan untuk mengubah data yang bersifat tidak baku ke bentuk yang sesuai aturan tata Bahasa Indonesia

d. *Stopword*
   Proses ini menghapus kata yang sering muncul namun tidak mempengaruhi akurasi dalam proses klasifikasi

e. *Stemming*
   Proses ini mengubah kata berimbuhan menjadi kata dasar sesuai aturan tata Bahasa Indonesia.

2.3 Ekstraksi Fitur Term-Frequency

Term-Frequency merupakan salah satu metode ekstraksi fitur yang mengubah kata menjadi representasi numerik yang dapat dimengerti oleh algoritma pemrosesan mesin [6]. Untuk menghitung *Term-Frequency* menggunakan persamaan berikut :

$$tf_{t,d} = \frac{jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{total\ jumlah\ kata\ dalam\ dokumen\ d} \quad (1)$$

2.4 Klasifikasi Naive Bayes

Metode *Multinomial Naive Bayes* merupakan pengembangan dari metode *Naive Bayes Classifier* (NBC) yang lebih unggul dalam pemrosesan data tekstual yang berdimensi lebih besar. Beberapa tahapan dalam proses klasifikasi data teks menggunakan metode *Multinomial Naive Bayes* adalah sebagai berikut [4]:

a. Menghitung probabilitas *class*
Perhitungan probabilitas class terhadap dokumen dapat dilakukan menggunakan persamaan berikut:

$$P(c) = \frac{N(c)}{N} \quad (2)$$

Keterangan:

$P(c)$ = probabilitas class terhadap dokumen
$N(c)$ = jumlah dokumen pada setiap class
N = jumlah seluruh dokumen

b. Membuat *term-document matrix*
Term-document matrix dibuat untuk menghitung jumlah kata pada semua dokumen, jumlah kata unik pada semua dokumen, dan jumlah kata pada setiap *class.*

c. Menghitung probabilitas kata unik
Perhitungan probabilitas kata unik untuk semua *class* dilakukan menggunakan persamaan berikut:

$$P(w|c) = \frac{count(w,c)+1}{count(c)+|V|} \quad (3)$$

Keterangan:

$P(w|c)$ = probabilitas munculnya kata unik *w* dalam dokumen yang termasuk dalam kelas *c*
$count(w,c)$ = jumlah kemunculan kata unik *w* dalam dokumen yang termasuk dalam kelas *c*
$count(c)$ = jumlah total kata dalam dokumen yang termasuk dalam kelas *c*
V = jumlah kata unik pada seluruh dokumen
| | = nilai mutlak

Pada persamaan di atas, digunakan teknik smooting Laplace untuk menghindari nilai probabilitas yang bernilai nol. Teknik smoothing Laplace dilakukan dengan menambahkan 1 pada jumlah kemunculan kata w dalam kelas c ($count(w,c)$) dan menambahkan jumlah kata unik (|V|) ke $count(c)$.

d. Menghitung probabilitas dokumen atau kalimat
Perhitungan probabilitas dokumen atau kalimat terhadap kelas dapat dilakukan menggunakan persamaan berikut:

$$P(c|d_{(n)}) = P(c)\ x\ \prod P(w|c) \quad (4)$$

Keterangan:

$P(c|d_{(n)})$ = probabilitas kalimat terhadap kelas
$P(c)$ = probabilitas kelas terhadap dokumen
$\prod$ = *product* (perkalian beruntun)
$P(w|c)$ = probabilitas kata terhadap kelas

2.5 Pengujian

Metode pengujian yang digunakan adalah metode *Confusion Matrix. Confusion matrix,* sebuah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam analisis sentimen atau masalah klasifikasi lainnya. *Confusion Matrix* digunakan untuk menghitung berbagai metrik evaluasi kinerja model seperti akurasi *(accuracy),* presisi *(precision), recall (sensitivity), atau* F1-Score.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1-Score = 2\frac{Precision \times Recall}{Precision+Recall} \quad (8)$$

3. Hasil dan Pembahasan

Berikut ini adalah hasil dari implementasi metode penelitian yang meliputi beberapa tahapan sebagai berikut :

3.1. Pengumpulan Data

Data yang digunakan merupakan ulasan dari produk Lampu 3D LED dari *Shopee.* Metode yang digunakan dalam proses pengumpulan data adalah metode *web scraping* menggunakan bahasa pemrograman bahasa Python. Ulasan yang dikumpulkan berjumlah 400 data dengan atribut *rating* dan *comment.* Data yang sudah terkumpul disimpan dalam bentuk CSV. Sebelum data digunakan untuk tahap berikutnya, data akan diimport dalam bentuk file Excel dengan format .xlsx. Berikut ini hasil dari pengumpulan data yang dapat dilihat pada Gambar 2.

	rating	comment
0	1	Barang yg datang tidak sesuai pesanan seler kurang teliti
1	1	Kualitas:buruk\nHarga:lumayan\n\nBaru kali ini saya sangat sangat kecewa bangettt lampu tidak menyala dan tidak berfungsi dengan baik demi allah kecewa banget ya walaupun harganya ga seberapa tpi tanggung jawabnya gaada sama sekali,cht juga ga di respon smoga allah yg membayarnya kalo bisa sebelum dikirim dicek dulu \nbarangnyaðŸ™□ðŸ□»
2	1	baru kali ini ngasi nilai bintang 1 sama penjual, udah patah ga idup pula lampunya. PARAH
3	1	Kualitas:lumayan bagus sh.\nHarga:murah\nDaya listrik:kurang jelas.\n\nGimana sh ini ngirim barang nya gay,, ko, ko ngash yg pecah alias belah. \n\nBisa engga y. Y di balikin atau tuker barang nys itu.
4	1	Tampilan:std\nPerforma:std\nKualitas:std\n\nKesel masalah pengiriman blgnya instant tau2 lama bgt ga dipick up expedisi blgnya blm diupdate buat kado jd telat
...	...	...
395	5	Walaupun pengemasan lama, tapi gpp. Barang sampai dgn baik, cakep nih sesuai foto.. Bakal order lagi buat kado lumayan
396	5	Buat kado semoga ikhen syukak lampunya bagus berkualitas dan mirah
397	5	Barangnya sudah sampai.. real picture.. kualitasnya bagus ðŸ□ðŸ□ðŸ□
398	5	Kualitas:bagus\nHarga:murahhh\nDaya listrikhemaat\n\nLucuuuu, terang banget lagi, luv dehhh pokonyaaaðŸ¥°ðŸ¥°ðŸ¥°ðŸ¥°
399	5	Tampilan:cakep\nPerforma:cakep\nKualitas:cakep\n\nKasih bintang 7 kalo ada ðŸ□ðŸ□

Gambar 2. Pengumpulan Data

3.2. Preprocessing

Tahap selanjutnya yaitu melakukan preprocessing untuk mengubah data yang bersifat tidak terstruktur menjadi data yang lebih tersrtuktur. Tahapan dalam preprocessing terdiri dari lima proses yang dijabarkan sebagai berikut:

a. *Cleaning Data*
Cleaning merupakan proses membersihkan data dari karakter yang tidak penting seperti tanda baca, angka, dan karakter lainnya. Pada proses ini juga dilakukan proses *Case Folding* yang mengubah huruf kapital menjadi huruf kecil. Proses *cleaning* pada salah satu data dapat dilihat pada Tabel 1.

Tabel 1. *Cleaning Data*

Data Awal	***Cleaning***
Barangnya sudah sampai. real picture. kualitasnya bagus ðŸ‘• ðŸ‘• ðŸ‘• ðŸ‘•	barangnya sudah sampai real picture kualitasnya bagus

b. *Tokenization*
Proses *tokenization* merupakan proses pemotongan string kalimat menjadi satuan token atau satuan kata. Proses *tokenization* dapat dilihat pada Tabel 2.

Tabel 2. *Tokenization*

Cleaning	***Tokenization***
barangnya sudah sampai real picture kualitasnya bagus	[barangnya, sudah, sampai, real, picture, kualitasnya, bagus]

c. *Normalization*
Proses ini dilakukan untuk mengubah data yang bersifat tidak baku ke bentuk yang sesuai aturan tata Bahasa Indonesia. Proses *normalization* dapat dilihat pada Tabel 3.

Tabel 3. *Normalization*

Tokenization	***Normalization***
[barangnya, sudah, sampai, real, picture, kualitasnya, bagus]	barangnya sudah sampai real picture kualitasnya bagus

d. *Stopword*
Proses ini menghapus kata yang sering muncul namun tidak mempengaruhi akurasi dalam proses klasifikasi. Proses *Stopword* dapat dilihat pada Tabel 4.

Tabel 4. *Stopword*

Normalization	***Stopword***
barangnya sudah sampai real picture kualitasnya bagus	barangnya real picture kualitasnya bagus

e. *Stemming*
Proses ini mengubah kata berimbuhan menjadi kata dasar sesuai aturan tata Bahasa Indonesia. Proses *Stemming* dapat dilihat pada Tabel 5.

Tabel 5. *Stemming*

Stopword	***Stemming***
barangnya sudah sampai real picture kualitasnya bagus	barang real picture kualitas bagus

3.3. Ekstraksi Fitur *Term-Frequency*

Ekstraksi fitur dilakukan untuk merepresentasikan data yang berbentuk teks menjadi data numerik sehingga dapat diolah pada proses klasifikasi. Dalam ekstraksi fitur *Term-Frequency*,

setiap kata dalam data akan dihitung jumlah kemunculannya dan digunakan sebagai fitur. Proses ekstraksi fitur ini dilakukan dengan menggunakan modul *CountVectorizer.*

3.4. Klasifikasi *Naive Bayes*

Setelah proses ekstraksi fitur selesai dilakukan, langkah selanjutnya yang dilakukan yaitu melakukan proses klasifikasi menggunakan metode *Multinomial Naive Bayes*. Proses klasifikasi dilakukan menggunakan *library MultinomialNB* dari pustaka *sklearn.naive_bayes*. Data dibagi dengan pembagian 80% sebagai data latih dan 20% sebagai data uji sehingga data latih berjumlah 320 data dan data uji berjumlah 80 data. Adapun hasil klasifikasi data uji dari proses klasifikasi yang dapat dilihat pada Tabel 6.

Tabel 6. Hasil Klasifikasi

Keterangan	Jumlah Prediksi
Prediksi Benar	69
Prediksi Salah	12

3.5. Pengujian

Proses pengujian merupakan tahap akhir yang bertujuan untuk mengevaluasi tingkat keberhasilan metode dalam mengklasifikasikan data. Pengujian dilakukan menggunakan tabel *confusion matrix*. Adapun hasil yang didapat dalam pengujian dapat dilihat pada tabel 7.

Tabel 7. *Confusion Matrix*

Keterangan	***Actual Positive***	***Actual Negative***
Predicted Positive	7	7
Predicted Negative	4	62

Dari *Confusion Matrix* di atas dapat diketahui nilai akurasi hasil klasifikasi sebesar 0,85. Setelah mendapatkan nilai *confusion matrix,* proses dilanjutkan dengan proses perhitungan nilai *precision, recall,* dan *F1-Score* yang dapat dilihat pada tabel 8.

Tabel 8. Nilai *precision, recall,* dan *F1-Score*

Keterangan	***Precision***	***Recall***	***F1-Score***
Positif	0,90	0,94	0,92
Negatif	0,64	0,50	0,56
Rata-rata	0,77	0,72	0,74

4. Kesimpulan

Dari penelitian yang telah dilakukan, dapat ditarik kesimpulan bahwa metode *Naive Bayes Classifier* dapat digunakan untuk klasifikasi data ulasan produk pada salah satu *e-commerce* yaitu *Shopee.* Berdasarkan penelitian di atas, didapatkan nilai akurasi sebesar 85%, nilai *precision* sebesar 77%, nilai *recall* sebesar 72%, dan nilai *f1-score* sebesar 74% dengan menggunakan data sebanyak 400 buah.

Daftar Pustaka

[1] Ahdiat, A. (2023, Mei 3). *5 E-Commerce dengan Pengunjung Terbanyak Kuartal I 2023*. Diambil kembali dari databoks: https://databoks.katadata.co.id/datapublish/2023/05/03/5-e-commerce-dengan-pengunjung-terbanyak-kuartal-i-2023

[2] Atmadja, B. R. (2022). Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata di Kabupaten Sukabumi Dengan Naive Bayes. *Jurnal Ilmiah Elektronika dan Komputer, Vol. 15, No. 2*, 371-382.
[3] Dedi Darwis, N. S. (2021). Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional. *Jurnal TEKNO KOMPAK, Vol. 15, No. 1*, 131-145.
[4] Febry Eka Purwiantono, A. A. (2020). Klasifikasi Sentimen SARA, Hoaks, dan Radikal pada Postingan Media Sosial menggunakan Algoritma Naive Bayes Multinomial Text. *Jurnal TEKNOKOMPAK, Vol. 14, No. 2*, 68-73.
[5] Febry Eka Purwiantono, A. A. (2020). Klasifikasi Sentimen SARA, Hoaks, dan Radikal pada Postingan Media Sosial Menggunakan Algoritma Naive Bayes Multinomial Text. *Jurnal TEKNOKOMPAK, Vol. 14, No. 2*, 68-73.
[6] Fika Hastaria Rachman, I. (2022). Pendekatan Data Science untuk Mengukur Empati Masyarakat terhadap Pandemi Menggunakan Analisis Sentimen dan Seleksi Fitur. *JEPIN (Jurnal Edukasi dan Penelitian Informatika) Vol. 8, No. 3*, 492-499.

Halaman ini sengaja dibiarkan kosong