

Analisis Sentimen Ulasan Traveloka Menggunakan Metode Naïve Bayes Classifier dan Information Gain

Kadek Yuni Suratri^{a1}, I Gede Santi Astawa^{a2}

Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Jalan Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia
¹yunisuratri@gmail.com
²santi.astawa@unud.ac.id

Abstract

In the increasingly rapid digital era, Traveloka is present as an online travel agency that makes it easier for users to order and plan trips. Reviews left by users can reflect the user's experience in using the platform. Indirectly, reviews can also reflect user satisfaction. Therefore, it is important to carry out sentiment analysis of existing reviews so that you can improve service quality. This research examines the performance of the Information Gain feature selection in classifying the sentiment of Traveloka application reviews using the Naïve Bayes method. The research results show that classification using the Naïve Bayes model obtained an accuracy of 83%, precision of 81%, and recall of 98%. Meanwhile, classification with feature selection obtained an accuracy of 79%, precision of 76%, and recall of 100%. This shows that the feature selection performance has not been able to increase the accuracy value.

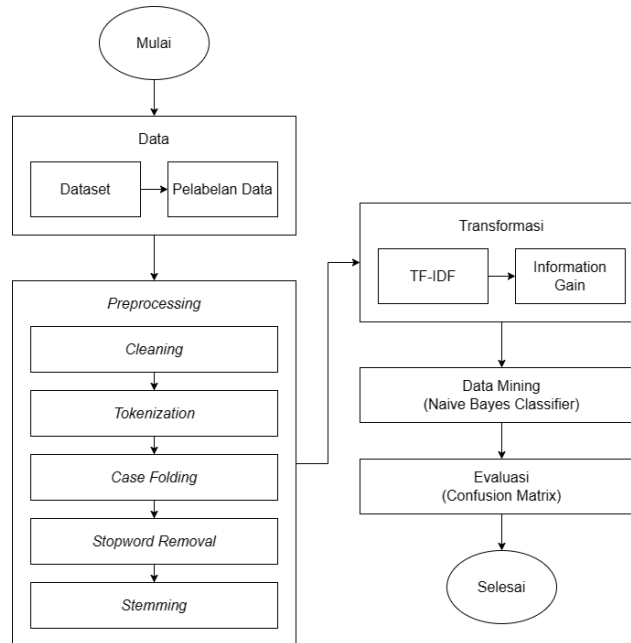
Keywords: Sentiment Analysis, Reviews, Traveloka, Naïve Bayes Classifier, TF-IDF, Information Gain

1. Pendahuluan

Di era digital yang berkembang semakin pesat, Traveloka hadir sebagai salah satu online travel agency yang memudahkan pengguna dalam pemesanan dan merencanakan perjalanan. Dilansir dari situs similarweb, total kunjungan pada situs Traveloka terhitung sampai bulan April 2024 telah mencapai sebesar 22.3 juta. Sebanyak 50 juta pengguna telah mengunduh aplikasi Traveloka pada Google Play Store. Hal ini menandakan suatu kepercayaan besar terhadap platform ini. Ulasan yang ditinggalkan oleh pengguna dapat mencerminkan pengalaman pengguna dalam menggunakan platform. Ulasan yang diberikan dapat berupa feedback, mulai dari pujian atas kemudahan dalam penggunaan platform hingga kritik terhadap segi layanan yang ditawarkan, masalah teknis. Secara tidak langsung, ulasan juga dapat mencerminkan kepuasan pengguna. Oleh karena itu, penting melakukan analisis sentimen terhadap ulasan yang ada sehingga dapat meningkatkan kualitas pelayanan. Dengan adanya peningkatan volume ulasan, diperlukan juga adanya suatu sistem untuk menganalisis sentimen secara otomatis. Analisis sentimen merupakan proses menganalisis teks dengan tujuan memperoleh informasi berupa sentimen baik positif, negatif ataupun netral di dalamnya. Analisis sentimen termasuk ke dalam salah satu bidang Natural Language Processing (NLP) yang berfokus pada identifikasi, ekstraksi, dan mengubah informasi subjektif dalam teks. Algoritma Naïve Bayes Classifier (NBC) merupakan salah satu algoritma yang populer dan sering terpilih dalam melakukan analisis sentimen. Hal ini karena algoritma NBC sederhana, mampu mengatasi missing value, mengelola atribut yang beragam, gangguan dalam data, memiliki kinerja yang baik sehingga cocok digunakan pada dataset dengan jumlah besar [1]. Dalam analisis sentimen, ekstraksi fitur memegang peranan penting sebagai jembatan antara data raw dan algoritma Machine Learning. Pada proses ini kompleksitas dan dimensi dari data teks yang sangat besar dapat berkurang, sehingga model dapat fokus pada elemen-elemen yang paling berpengaruh terhadap sentimen. Selain itu, ekstraksi fitur dapat membantu dalam mengeliminasi noise yang dapat merusak akurasi prediksi. Pada penelitian ini akan menguji performa seleksi fitur Information Gain (IG) dalam mengoptimalkan performa algoritma NBC.

2. Metode Penelitian

Penelitian ini mengimplementasikan metode Knowledge Discovery in Database (KDD) yang meliputi beberapa tahapan seperti pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola ataupun hubungan pada dataset yang berukuran besar [2].



Gambar 1. Tahap Penelitian

Penelitian ini diawali dengan tahap mengumpulkan data. Data yang telah dikumpulkan akan diberi label. Selanjutnya akan dilakukan tahap preprocessing teks. Kemudian data yang telah bersih akan masuk ke tahap transformasi. Tahapan transformasi diawali dengan penghitungan ekstraksi fitur menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dan dilanjutkan ke proses seleksi fitur dengan Information Gain (IG). Berikutnya masuk ke tahap klasifikasi data dengan metode Naive Bayes Classifier. Terakhir merupakan tahap evaluasi dengan confusion matrix. Hal ini dilakukan untuk mengukur validitas dari tahap klasifikasi.

2.1. Dataset

reviewid	userName	userImage	content	score	thumbsUp	thumbsDown	reviewCreate at	replyContent	repliedAt	appVersion
0_9d55103c-b1	Muh Supriya	https://play-	Ok	5	0	0,1861111	22/10/23 5.10	Hi, thank you	22/10/23 5.12	0,1861111
1_a51b3420-e1	Sry Hary	https://play-	Lebih Memui	5	0	0,1861111	22/10/23 5.06	Hi Kak, senan	22/10/23 5.12	0,1861111
2_f47e4b05-85	Sligt Mamya	https://play-	Selalu pakal	5	0	0,1861111	22/10/23 4.59	Halo Kak, teri	22/10/23 5.12	0,1861111
3_0dc69506-fe	Awawy Alawj	https://play-	memudahkai	5	0	0,1861111	22/10/23 4.55	Hi Kak, senan	22/10/23 5.12	0,1861111
4_10a64a1b-ct	Uchu Abm	https://play-	Up	5	0	0,1819444	22/10/23 4.51	Thank you fo	22/10/23 5.12	0,1819444
5_3b2b2d7a-5	Sumiati	https://play-	The best app	5	0	0,1861111	22/10/23 4.50	Hi, thank you	22/10/23 5.12	0,1861111
6_1d5d15d5-a1	Ryan Andria	https://play-	Aplikasinya b	5	0	0,1861111	22/10/23 4.38	Halo Kak, teri	22/10/23 4.42	0,1861111
7_c39c5886-e7	Isah Ruhayu	https://play-	Untuk cari tik	5	0	0,1861111	22/10/23 4.33	Hi Kak, senan	22/10/23 4.42	0,1861111
8_1a8b72d6-d1	Asih Karyati	https://play-	Mempermud	5	0	0,1861111	22/10/23 4.31	Hi Kak, terim	22/10/23 4.42	0,1861111
9_f8c90ded-85	Ninda Khana	https://play-	Aplikasi anda	5	0	0,1736111	22/10/23 4.31	Halo Kak, teri	22/10/23 4.42	0,1736111
10_52ad09c1-b1	William They	https://play-	Mau telepon	1	0	0,1861111	22/10/23 3.06	Hi William, #	22/10/23 3.23	0,1861111
11_5874c218-az	Jonathan Cric	https://play-	Traveloka ma	5	0	0,1861111	22/10/23 2.39	Halo Kak, teri	22/10/23 3.42	0,1861111
12_a88c832d-f5	Lilis	https://play-	Penanganan	1	0	0,1861111	22/10/23 2.37	Hai Lilis, moh	20/10/23 8.51	0,1861111
13_f192bc00-55	yanti okedeh	https://play-	ok	5	0	0,1694444	22/10/23 0.55	Hi, thank you	22/10/23 1.12	0,1694444
14_22c69269-b1	Satriojagat8	https://play-	Bagus tingka	5	0	0,1861111	22/10/23 0.50	Hi Kak, senan	22/10/23 1.12	0,1861111
15_c44546f0-65	shufatun mu	https://play-	bikin gampar	5	0	0,1861111	22/10/23 0.10	Halo Kak, teri	22/10/23 0.12	0,1861111
16_7d4b7a6f-e1	Dian Wiantok	https://play-	Kurang puas	1	0	0,1861111	21/10/23 23.11	Hai Dian, moi	21/10/23 23.23	0,1861111
17_0024a58e-e1	Uud Al_mun	https://play-	Terima kasih	5	0	0,1861111	21/10/23 23.04	Halo Kak, teri	21/10/23 23.12	0,1861111
18_d6174614-a1	Muhamad Isr	https://play-	Pas di peruku	1	0	0,1861111	21/10/23 22.49	Hai Muhama	21/10/23 23.19	0,1861111
19_dd2b0dc9-31	Saleh Suhenc	https://play-	Alhamdulillah	5	0	0,1861111	21/10/23 18.04	Halo Kak, teri	21/10/23 18.12	0,1861111
20_523fec84-15	Tomri Siregar	https://play-	Baddd	1	0	0,1861111	21/10/23 16.52	Hi Tomri, We	21/10/23 19.15	0,1861111
21_45f06493-11	Rindu Duma	https://play-	bagus	5	0	0,1798611	21/10/23 15.52	Hi Kak, terim	21/10/23 16.12	0,1798611
22_57f66b14-94	Awak Droe	https://play-	php jgn mau	1	0	0,1861111	21/10/23 15.08	Hi Kak, senan	21/10/23 11.12	0,1861111
23_7613258c-71	Daddy Rashid	https://play-	Bagus semog	5	0	0,1861111	21/10/23 14.51	Halo Kak, teri	21/10/23 15.12	0,1861111
24_930d79fa-d1	Frans Yonq	https://play-	Mudah dan g	5	0	0,1861111	21/10/23 13.51	Halo Kak, teri	21/10/23 14.12	0,1861111
25_81866496-84	Aguss Fajar	https://play-	top	5	0	0,1854167	21/10/23 12.16	Hi, we are so	21/10/23 12.42	0,1854167
26_363c8210-c1	Dibyio Utomc	https://play-	Aplikasi paya	1	4	0,1854167	21/10/23 12.08	Hai Dibyio, m	13/10/23 16.30	0,1854167

Gambar 2. Dataset

Penelitian ini menggunakan dataset yang telah tersedia di situs Kaggle. Proses pengambilan data dilakukan dengan mengunduh secara langsung dari situs. Total dataset berjumlah 177.159, namun pada penelitian hanya digunakan sebanyak 1.501 data. Pemberian label pada dataset dilakukan dengan bahasa pemrograman Python dan menjadikan score sebagai indikator pelabelan. Ulasan akan dikelompokkan ke dalam 3 kelas sentimen yaitu sentiment positif, negatif dan netral.

2.2. Preprocessing Text

Teks preprocessing merupakan proses yang dilakukan sebelum data teks diolah menggunakan algoritma NLP dengan tujuan pembersihan dan penyusunan ulang data teks agar terstruktur, sehingga lebih mudah diolah oleh algoritma. Tahap pertama dari preprocessing yaitu cleaning, membersihkan teks dari segala elemen yang tidak relevan seperti menghapus URL, karakter khusus, angka, dan sebagainya. Dilanjutkan pada tahap kedua yaitu tokenization, memecah kalimat menjadi sebuah kata atau yang disebut token. Dilanjutkan pada tahap ketiga yaitu case folding, mengubah keseluruhan huruf menjadi huruf kecil. Tahap keempat yaitu stopword removal, menghapus kata yang tidak diperlukan seperti kata konjungsi. Tahap terakhir yaitu stemming, mengubah kata ke bentuk dasar dengan cara menghilangkan imbuhan dari kata tersebut.

2.3. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF merupakan algoritma dari ekstraksi fitur untuk mengukur seberapa penting sebuah kata dalam dokumen. Nilai TF-IDF didapat dengan mengalikan nilai dari Term Frequency (TF) dan Inverse Document Frequency (IDF). TF menghitung seberapa sering kemunculan suatu kata pada sebuah dokumen sedangkan IDF menghitung seberapa jarang/tidak sering munculnya suatu kata di dokumen lain [3]. Adapun representasi sistematis untuk menghitung pembobotan setiap token t di dokumen d sebagai berikut:

$$tfidf_t = f_{t,d} \times \log \frac{N}{df_t} \quad (1)$$

$tfidf_t$ merupakan bobot dari kata ke- t , $f_{t,d}$ merupakan kemunculan kata ke- t dalam dokumen ke- d , N merupakan total dari dokumen, dan df_t merupakan banyaknya dokumen yang mengandung kata- t .

2.4. Information Gain

Information gain (IG) merupakan teknik yang digunakan dalam seleksi fitur dengan menggunakan metode penilaian untuk fitur kategori dan juga untuk pembobotan atribut yang berkelanjutan (kontinu) yang kemudian diubah menjadi nilai diskrit menggunakan entropi maksimal [4]. Suatu entropi digunakan untuk mengukur jumlah informasi yang diperlukan untuk mengkodekan suatu kelas [5]. IG dari suatu fitur diukur dengan menghitung jumlah informasi yang diperoleh dari prediksi kategori dengan mempertimbangkan keberadaan atau ketiadaan fitur tersebut dalam dokumen. Adapun representasi sistematisnya sebagai berikut:

$$InfoGain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (2)$$

$$Entropy(S) = - \sum \frac{|S_i|}{S} \log \frac{S_i}{S} \quad (3)$$

S merupakan jumlah seluruh fitur, A merupakan kategori, S_v merupakan jumlah sampel untuk nilai v , v merupakan nilai yang mungkin untuk kategori A , S_i merupakan fitur ke i dan $Value(A)$ merupakan himpunan nilai-nilai yang mungkin untuk kategori A . Fitur yang dipilih merupakan fitur yang memiliki IG yang berbeda dari nol dan lebih besar dari nilai threshold tertentu. Konsep dibalik penggunaan IG untuk menentukan fitur yang paling signifikan dalam menjelaskan kategori.

2.5. Algoritma Naïve Bayes Classifier

Algoritma Naïve Bayes Classifier (NBC) merupakan sebuah model klasifikasi dengan menggunakan prinsip Naïve Bayes dalam memprediksi kelas dari suatu data berdasarkan fitur-fiturnya. Naive Bayes Classifier sendiri merupakan metode pembelajaran probabilistik di mana setiap kata memiliki probabilitas kemunculan yang independen. Ini berarti bahwa nilai atribut kategori tidak saling mempengaruhi atau dipengaruhi oleh nilai atribut lainnya [6].

2.6. Evaluasi

Tahap terakhir merupakan tahap evaluasi dari performa model klasifikasi dengan confusion matrix. Confusion matrix merupakan sebuah tabel dengan informasi di dalamnya tentang jumlah prediksi keseluruhan benar dan salah yang dilakukan oleh model untuk setiap kelas dari data. Hal yang akan dihitung adalah nilai accuracy, recall dan precision.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (5)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (6)$$

TP merupakan jumlah data positif yang diprediksikan benar sebagai nilai positif, FN merupakan jumlah data positif yang salah diklasifikasikan sebagai nilai negatif, FP merupakan jumlah data negatif yang salah diklasifikasikan sebagai nilai positif, dan TN merupakan jumlah data negatif yang diprediksi benar sebagai nilai negatif.

3. Hasil dan Diskusi

3.1. Preprocessing

Dilakukan tahap preprocessing pada data sebelum memasuki tahap berikutnya. Pada tahap preprocessing, data akan dibersihkan dengan menghilangkan noise melalui penghapusan URL, simbol, angka, dan lain-lain. Dataset yang akan di preprocessing berjumlah 1.232 karena telah melalui proses penghilangan redundansi.

Tabel 1. Hasil Preprocessing

Tahap preprocessing	Hasil
Dataset	Selalu pakai Traveloka untuk pembelian tiket dan hotel kemana pun... 🙌
Cleaning	Selalu pakai Traveloka untuk pembelian tiket dan hotel kemana pun
Tokenization dan Case Folding	selalu, pakai,traveloka,untuk,pembelian,tiket,dan,hotel,kemana,pun
Stopword Removal	pakai,traveloka,pembelian,tiket,hotel,kemana
Stemming	pakai traveloka beli tiket hotel mana

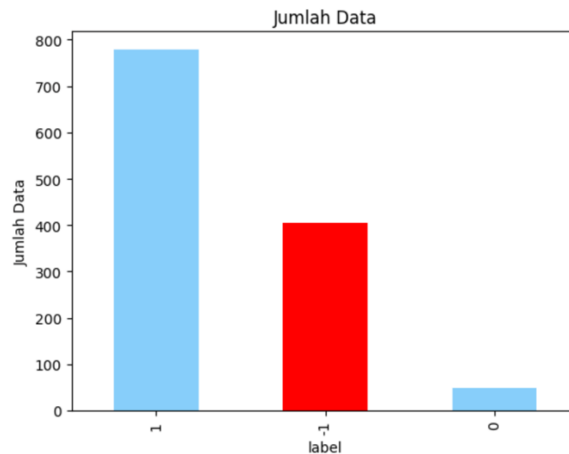
3.1.1. Labeling

Pada labeling, akan dilakukan pengklasifikasian sentimen ke dalam 3 kelas kata yaitu sentimen positif, negatif, dan netral [7].

```
#labeling
def pelabelan(skor):
    if skor < 3:
        return '-1'
    elif skor == 3:
        return '0'
    elif skor > 3 :
        return '1'
data['label'] = data['skor'].apply(pelabelan)
data.head(10)
```

Gambar 3. Labeling

Pada gambar 3, labeling dilakukan berdasarkan dengan score atau rating yang diberikan oleh pengguna. Apabila skor yang diberikan bernilai < 3 (lebih kecil dari 3), maka sentimen akan diklasifikasikan menjadi sentimen negatif. Skor == 3 merupakan sentimen netral, dan skor > 3 (kurang dari 3) merupakan sentimen positif. Didapat sebanyak 405 data dengan label negatif, 779 data dengan label positif, dan 48 data dengan label netral. Gambar 4 merupakan visualisasi persebaran label menggunakan grafik batang.



Gambar 4. Persebaran Label

Tabel 2. Contoh Data Label

Data	Label
aplikasi eror	Negatif
aplikasi curang rubah tanggal sistem buruk	Negatif
aplikasi digital mesan pesawat hotel murah	Positif
aplikasi nya keren bantu cari hotel tiket pesawat	Positif
kk sy pesan hotel tp udah sy batal knp sy ttp bayar ya	Netral
Aplikasi lot	Netral

3.1.2. Pembobotan Kata

Pada tahap pembobotan kata menggunakan TF-IDF (Term Frequency – Inverse Document Frequency). Sebelum dilakukan TF-IDF, dataset akan dibagi menjadi data train dan data test dengan rasio perbandingan berturut-turut 80:20. Jumlah data train sebanyak 985 dan jumlah data test sebanyak 247.

Gambar 4 merupakan representasi sparse matrix dari hasil transformasi teks menggunakan 'TfidfVectorizer'. Kolom pertama menunjukkan dokumen, kolom kedua menunjukkan indeks dan kolom ketiga menunjukkan nilai TF-IDF. Misal "(0, 1788) 0.29490689096159084" berarti dalam dokumen ke-0 atau pertama, kata dengan indeks 1788 dalam vocabulary memiliki nilai TF-IDF sebesar 0.29490689096159084.

(0, 1788)	0.29490689096159084
(0, 1749)	0.28344098627145816
(0, 1547)	0.29490689096159084
(0, 1488)	0.5347453281386314
(0, 1416)	0.3356911165806572
(0, 1015)	0.167845552893286
(0, 981)	0.29490689096159084
(0, 919)	0.2485538353615331
(0, 627)	0.3881585065750939
(0, 10)	0.27888054464483703
(1, 1924)	0.5815909619844721
(1, 212)	0.5687748376247432
(1, 98)	0.5815909619844721
(2, 2844)	0.28249658857617826
(2, 1876)	0.2528263096541089
(2, 1828)	0.2713844046415995
(2, 1817)	0.3723649277495937
(2, 1612)	0.3895386972067483
(2, 1344)	0.17275794634838468
(2, 1086)	0.274764849617055
(2, 493)	0.3723649277495937
(2, 374)	0.254428128262288
(2, 189)	0.38116120045818
(2, 130)	0.18489642778828288
(3, 2145)	0.246524935640334
:	:
(981, 707)	0.2908267443884993
(981, 329)	0.23189798438653218
(981, 297)	0.22555187859793405
(981, 265)	0.2888569167077537
(981, 160)	0.12655635801122353
(981, 93)	0.18622078114242384
(982, 1936)	0.13635950099188524
(982, 1758)	0.35912772665756293
(982, 1466)	0.18445498687843233
(982, 1389)	0.20960316730750028
(982, 1200)	0.325597767654457
(982, 1079)	0.28637823851938774
(982, 983)	0.35912772665756293
(982, 889)	0.26858209361793334
(982, 489)	0.33951395799002826
(982, 288)	0.21543836594729282
(982, 246)	0.35912772665756293
(982, 228)	0.28127357282782994
(982, 93)	0.1247974156244584
(983, 1329)	1.0
(984, 1344)	0.2974996991871526
(984, 1243)	0.6412350718848003
(984, 919)	0.479526878342742
(984, 671)	0.36673836246529307
(984, 29)	0.3685567997344024

Gambar 4. Hasil Pembobotan Kata

3.1.3. Seleksi Fitur

Proses seleksi fitur menggunakan algoritma Information Gain (IG). Dengan adanya seleksi fitur, fitur yang tidak relevan akan dikurangi. Sebelum dilakukan seleksi fitur, jumlah fitur sebanyak 2.158. Dengan menggunakan IG threshold 0.0008 sebanyak 690 fitur berhasil dihilangkan sehingga menyisakan 1.468 fitur.

Jumlah fitur sebelum seleksi fitur: 2158
 Jumlah fitur setelah seleksi fitur: 1468

Gambar 5. Hasil Seleksi Fitur

3.1.4. Evaluasi

Tahapan klasifikasi menggunakan bantuan model dari algoritma Naïve Bayes yaitu Multinomial Naïve Bayes (NB). Setelah tahap klasifikasi, tahap evaluasi dilakukan dengan menggunakan confusion matrix. Hasil berupa confusion matrix terhadap model yang telah dibuat dapat dilihat pada gambar 6 dan gambar 7.

Confusion Matrix:

$$\begin{bmatrix} 51 & 0 & 30 \\ 4 & 0 & 6 \\ 3 & 0 & 153 \end{bmatrix}$$

Gambar 6. Confusion Matrix Naïve Bayes

Confusion Matrix:

$$\begin{bmatrix} 39 & 0 & 42 \\ 3 & 0 & 7 \\ 0 & 0 & 156 \end{bmatrix}$$

Gambar 7. Confusion Matrix Naïve Bayes + IG

Berdasarkan gambar 6, diperoleh accuracy sebesar 83%, precision sebesar 81%, dan recall sebesar 98%. Hasil klasifikasi perolehan dengan model Naïve Bayes dapat dilihat pada gambar 8.

Classification Report:				
	precision	recall	f1-score	support
-1	0.88	0.63	0.73	81
0	0.00	0.00	0.00	10
1	0.81	0.98	0.89	156
accuracy			0.83	247
macro avg	0.56	0.54	0.54	247
weighted avg	0.80	0.83	0.80	247

Gambar 8. Klasifikasi dengan Naïve Bayes

Berdasarkan gambar 7, diperoleh accuracy sebesar 79%, precision sebesar 76%, dan recall sebesar 100%. Hasil klasifikasi perolehan dengan model Naïve Bayes dan seleksi fitur IG dapat dilihat pada gambar 9.

Classification Report:				
	precision	recall	f1-score	support
-1	0.93	0.48	0.63	81
0	0.00	0.00	0.00	10
1	0.76	1.00	0.86	156
accuracy			0.79	247
macro avg	0.56	0.49	0.50	247
weighted avg	0.79	0.79	0.75	247

Gambar 9. Klasifikasi Naïve Bayes dan Information Gain

Pada tabel 3 merupakan perbandingan performa model antara Naïve Bayes dan Naïve Bayes + IG berdasarkan akurasi.

Tabel 3. Perbandingan Performa Model Berdasarkan Accuracy

Metode	Akurasi
Naïve Bayes	83%
Naïve Bayes + IG	79%

Berdasarkan tabel perbandingan hasil performa, didapat bahwa performa kinerja dari model Naive Bayes lebih besar dibandingkan dengan performa kinerja dengan menggunakan tambahan seleksi fitur.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan terhadap penggunaan seleksi fitur dengan menggunakan metode Naïve Bayes, performa seleksi fitur belum cukup baik jika dibandingkan dengan tidak menggunakan seleksi fitur. Diperoleh nilai akurasi sebesar 79%, dimana seleksi fitur belum mampu meningkatkan akurasi klasifikasi. Adapun hal yang dapat diperbaiki untuk

penelitian kedepannya yaitu pada pelabelan dapat meminta bantuan ahli untuk menganalisis supaya lebih akurat, dan menggunakan dataset dalam jumlah persebaran label yangimbang.

Daftar Pustaka

- [1] Q. A'yuniyah et al., "Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik," *J. Sist. Komput. dan Inform.*, vol. 4, no. 1, p. 72, 2022, doi: 10.30865/json.v4i1.4781.
- [2] D. S. O. Panggabean, E. Buulolo, and N. Silalahi, "Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, p. 56, 2020, doi: 10.30865/jurikom.v7i1.1947.
- [3] I. Syahrohim, S. D. Saputra, and R. W. Saputra, "Pilpres 2024 Twitter Menggunakan," vol. 12, no. 2, 2024.
- [4] I. Maulida, A. Suyatno, and H. R. Hatta, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," *J. SIFO Mikroskil*, vol. 17, no. 2, pp. 249–258, 2016, doi: 10.55601/jsm.v17i2.379.
- [5] D. Abadi, "Perbandingan algoritme feature selection information gain dan symmetrical uncertainty pada data ketahanan pangan," 2013.
- [6] A. Isnanda, Y. Umaidah, and J. H. Jaman, "Implementasi Naïve Bayes Classifier Dan Information Gain Pada Analisis Sentimen Penggunaan E-Wallet Saat Pandemi," *J. Teknol. Inform. dan Komput.*, vol. 7, no. 2, pp. 144–153, 2021, doi: 10.37012/jtik.v7i2.648.
- [7] N. R. Siahaan, R. Y. Tiffany, and S. R. E. Sinaga, "Analisis Sentimen Ulasan Aplikasi Media Sosial Whatsapp Menggunakan Metode Naive Bayes Classifier," *J. Ilm. Betrik*, vol. 14, no. 02, pp. 343–354, 2023, [Online]. Available: <https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/104%0Ahttps://ejournal.pppmitpa.or.id/index.php/betrik/article/download/104/76>