Volume 11 Nomor 3, 2025, 175 -183
ISSN: 2460-1799 P-ISSN: 1432-602743

Klasifikasi Ulasan Pengguna Tiket Pesawat Online dengan Penanganan Ketidakseimbangan Data Menggunakan SMOTE dengan Machine Learning

Rahayun Amrullah Husaini[1*], Gede Yogi Pratama[2] Azral Satrani[2]

younk@universitasbumigora.ac.id [1,*], gedeyogipratama@universitasbumigora.ac.id [2], azral@universitasbumigora.ac.id [3]

[1] Teknologi Informasi; Fakultas Teknik, Universitas Bumigora; Jalan Ismail Marzuki No.22, Cilinaya, Cakranegara, Kota Mataram, Nusa Tenggara Barat, Telp: (0370)-638369,
[2] Ilmu Komputer; Fakultas Teknik Universitas Bumigora; Jalan Ismail Marzuki No.22, Cilinaya, Cakranegara, Kota Mataram, Nusa Tenggara Barat, Telp: (0370)-638369,
[3] Ilmu Komputer; Fakultas Teknik Universitas Bumigora; Jalan Ismail Marzuki No.22, Cilinaya, Cakranegara, Kota Mataram, Nusa Tenggara Barat, Telp: (0370)-638369,

* e-mail: younk@universitasbumigora.ac.id

Diterima: 25/12/2025 ; Review: 27/12/2025; Disetujui: 29/12/2025

ABSTRAK *– Pandemi COVID-19 telah memengaruhi kebiasaan masyarakat dalam melakukan perjalanan udara dan mendorong peningkatan pemesanan tiket secara daring. Penelitian ini bertujuan untuk menganalisis sentimen ulasan pembelian tiket pesawat secara daring menggunakan metode Support Vector Machine dan K-Nearest Neighbor dengan menggunakan teknik SMOTE. Penelitian ini dilakukan melalui pengumpulan data ulasan dari situs Tiket.com, dilanjutkan dengan tahap pra-pemrosesan, pembobotan menggunakan TF-IDF, serta klasifikasi menggunakan kedua metode tersebut. Hasil penelitian menunjukkan bahwa metode Support Vector Machine memperoleh akurasi sebesar 51 persen, sedangkan K-Nearest Neighbor mencapai akurasi 55 persen setelah penerapan teknik penyeimbangan data. Simpulan dari penelitian ini adalah bahwa kedua metode mampu mengklasifikasikan sentimen ulasan pengguna secara efektif dan dapat digunakan untuk membantu penyedia layanan dalam meningkatkan kualitas pelayanan serta memahami kebutuhan pelanggan.*

Kata Kunci *– Analisis Sentimen, Tiket Pesawat Online, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), SMOTE-Tomek Links.*

Classification of Online Airline Ticket User Reviews with Imbalanced Data Handling Using SMOTE and Machine Learning

ABSTRACT *– The COVID-19 pandemic affected public habits in air travel and increased the use of online ticket booking platforms. This study aimed to analyze sentiment in online flight ticket purchase reviews using the Support Vector Machine and K-Nearest Neighbor methods. The research was conducted by collecting user review data from the Tiket.com website, followed by preprocessing, term weighting using TF-IDF, and classification using both methods. The results show that the Support Vector Machine method achieves an accuracy of 51 percent, while the K-Nearest Neighbor method reaches 55 percent after applying data balancing techniques. This study concludes that both methods are effective in classifying user sentiment and can assist service providers in improving service quality and understanding customer needs..*

Keywords *– Sentiment Analysis, Online Flight Tickets, Support Vector Machine, K-Nearest Neighbor, Data Balancing.*

This is an open access article under the CC BY-SA license

1. PENDAHULUAN

Perjalanan udara telah menjadi kebutuhan umum masyarakat modern untuk berbagai keperluan seperti wisata, bisnis, dan kunjungan keluarga. Namun, pandemi COVID-19 mengubah kebiasaan ini secara signifikan, di mana masyarakat menjadi lebih waspada dalam bepergian menggunakan pesawat [1]. Maskapai penerbangan pun menerapkan berbagai kebijakan seperti physical distancing dan promo diskon untuk menarik kembali minat pelanggan. Di saat yang sama, tren pemesanan tiket secara daring melalui platform seperti Tiket.com, Traveloka, dan Booking.com semakin meningkat, memberikan kemudahan dan fleksibilitas bagi pengguna [2].

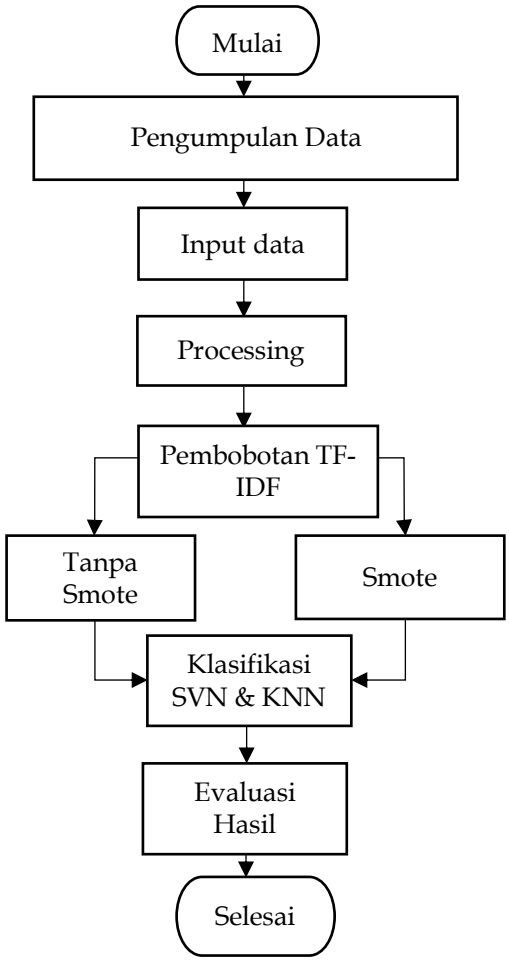
Di tengah kemudahan ini, ulasan pelanggan terhadap layanan pemesanan tiket pesawat menjadi sumber data penting yang dapat dianalisis untuk memahami kepuasan pelanggan [3]. Dalam konteks ini, analisis sentimen menjadi pendekatan yang tepat untuk menilai apakah ulasan yang diberikan bersifat positif atau negative [4]. Teknik ini menjadi bagian dari metode data mining yang mampu mengolah data teks dalam jumlah besar untuk menggali informasi yang berguna. Salah satu manfaat utamanya adalah membantu penyedia layanan memahami kebutuhan konsumen, meningkatkan layanan, serta menyusun strategi promosi berdasarkan preferensi pasar [5].

Penelitian-penelitian terdahulu telah menunjukkan efektivitas algoritma klasifikasi seperti *Support Vector Machine* dan K-Nearest Neighbor dalam menganalisis data ulasan [6]. Kedua algoritma ini terbukti memiliki akurasi tinggi dalam proses klasifikasi teks [6]. Oleh karena itu, penelitian ini bertujuan untuk menerapkan kedua metode tersebut dalam analisis sentimen pada ulasan pembelian tiket pesawat secara daring. Kebaruan dalam penelitian ini terletak pada penggunaan gabungan SVM dan KNN terhadap data ulasan dari situs Tiket.com yang diproses dengan teknik penyeimbangan data SMOTE-Tomek Links [7].

Permasalahan utama yang ingin dijawab dalam penelitian ini adalah bagaimana menerapkan metode Support Vector Machine dan K-Nearest Neighbor untuk analisis sentimen, serta bagaimana membandingkan performa kedua metode tersebut. Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem klasifikasi ulasan pelanggan, yang bermanfaat bagi penyedia layanan tiket online maupun masyarakat umum yang mencari informasi terpercaya sebelum melakukan pembelian.

2. METODE

Penelitian ini dilaksanakan dengan pendekatan eksperimental yang sistematis dan terstruktur, terdiri dari beberapa tahapan penting yang saling berkaitan. Tahapan pertama adalah pengumpulan data, yaitu mengumpulkan ulasan pengguna dari berbagai sumber yang relevan sebagai dasar analisis sentimen. Selanjutnya dilakukan preprocessing data, yang mencakup pembersihan teks seperti penghapusan tanda baca, stopword, dan transformasi menjadi huruf kecil untuk memastikan konsistensi data. Setelah itu, diterapkan proses pembobotan menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency), yang bertujuan untuk memberikan bobot terhadap kata-kata penting dalam ulasan. Agar data menjadi seimbang, dilakukan proses penyeimbangan data menggunakan teknik SMOTE-Tomek Links, yang membantu meningkatkan performa klasifikasi. Kemudian dilakukan penerapan metode klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN) untuk mengklasifikasikan sentimen. Terakhir, hasil klasifikasi dievaluasi menggunakan confusion matrix untuk mengukur akurasi, presisi, recall, dan F1-score, sebagaimana ditunjukkan pada Gambar 1.



Gambar. 1 Alur Penelitian

Penelitian ini memiliki tahapan penting yaitu pengumpulan data atau scrapping data, preprocessing data, pembobotan TF-IDF, Penyeimbangan data menggunakan SMOTE-Tomek Links, dan mengklasifikasikan topik dengan menggunakan metode SVM dan KNN untuk mendapatkan tingkat akurasi dari penelitian ini. Pengumpulan data dilakukan melalui teknik scraping pada situs Tiket.com, dengan mengambil ulasan pelanggan yang berkaitan dengan pembelian tiket pesawat. Dataset yang diperoleh berjumlah 500 komentar yang disertai nilai rating dan diklasifikasikan ke dalam dua kategori sentimen: positif dan negative [8]. Tahapan preprocessing meliputi proses cleaning data (menghapus karakter tidak penting), case folding (mengubah huruf menjadi kecil), tokenizing (memecah kalimat menjadi kata), stopword removal (menghapus kata tidak bermakna seperti "dan", "atau"), serta stemming (mengembalikan kata ke bentuk dasarnya). Proses ini bertujuan untuk menyederhanakan data sebelum dilakukan pembobotan [9]. Pembobotan kata dilakukan menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) untuk mengukur tingkat kepentingan suatu kata dalam dokumen. Setelah pembobotan, dataset disesuaikan dengan teknik SMOTE-Tomek Links guna mengatasi ketidakseimbangan jumlah data pada masing-masing kelas sentiment [10]. Selanjutnya, dua algoritma klasifikasi yaitu Support Vector Machine dan K-Nearest Neighbor diterapkan untuk melakukan klasifikasi terhadap data ulasan yang telah diproses. Algoritma SVM bekerja dengan membentuk hyperplane optimal untuk memisahkan dua kelas data, sedangkan KNN mengklasifikasikan data berdasarkan kedekatan jarak dengan data latih [10]. Terakhir, evaluasi model dilakukan menggunakan confusion matrix untuk mengukur akurasi, presisi, dan recall dari masing-masing metode. Perbandingan performa metode tanpa dan dengan SMOTE-Tomek Links juga dilakukan untuk mengetahui pengaruh teknik penyeimbangan terhadap hasil klasifikasi [11].

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari implementasi dan pengujian metode yang telah dijelaskan pada bab sebelumnya, serta analisis terhadap temuan yang diperoleh selama proses penelitian. Penelitian ini dilakukan dengan tujuan untuk menganalisis sentimen pengguna terhadap layanan pemesanan tiket pesawat secara online dengan menggunakan algoritma klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN). Setiap tahapan penelitian dijabarkan secara rinci mulai dari proses persiapan data, transformasi data, penerapan metode klasifikasi, hingga evaluasi hasil menggunakan confusion matrix.

Pembahasan pada bab ini difokuskan pada bagaimana metode yang digunakan mampu mengolah data ulasan pengguna menjadi informasi sentimen yang bermakna. Selain itu, dibandingkan pula kinerja masing-masing algoritma untuk melihat metode mana yang lebih akurat dalam mengklasifikasikan sentimen. Evaluasi dilakukan berdasarkan parameter seperti akurasi, presisi, recall, dan F1-score. Melalui analisis ini, diharapkan diperoleh pemahaman yang komprehensif mengenai efektivitas pendekatan yang digunakan dalam mengklasifikasikan sentimen pengguna secara otomatis dan sistematis.

Persiapan Data

Dalam persiapan data, langkah-langkah yang dilakukan melibatkan penggunaan aplikasi Google Colabs dan dengan menggunakan bahasa pemrograman Python [12]. Data dalam penelitian ini diperoleh dari scrapping data yang didapatkan melalui ulasan pengguna dari situs . jumlah data yang diambil berjumlah 500 data. Hasil scrapping data ditunjukkan pada Tabel 1.

Tabel. 1. Scrapping Data

No	at	userName	Ulasan	score
0	2023-07-27	Yandi Li	Lengkap, saat refund tidak	5
1	2023-03-01	Vera	Harga gak II jauh beda dgn apps	2
2	2023-08-03	Tiara Amelia	Respon lama., mau booking	1
3	2023-06-05	yugi engineer	Sangal mengecewakan	1
4	2023-07-20	Fajar Rizkika	Setelah update. login sulit	3
….	…..	…..	….	…
495	2021-01-23	Yanto	Reaktif saat rapid test	1
496	2020-05-02	Ade Eiwin	saya udah mengajukan	1
497	2018-09-13	Pengguna Gongle	Dari abis update malah ngk bisa	1
498	2021-01-01	Del lingga P K 5 P	Kecewa. free rapid antigen yg	1
499	2022-11-01	Gosnertin Sianturi	Sangat mengecewakan.	1

500 rows x 4

Setelah berhasil melakukan scraping data, langkah selanjutnya adalah menyortir data yang akan ditampilkan hanya "ulasan" dan "score" yang digunakan dalam penelitian [13]. Hasil dari data yang telah disaring ini kemudian ditampilkan pada Tabel 2.

Tabel. 2. Menampilkan Data

No	Ulasan	Score
0	Lengkap, saat refund tidak lama seria tidak me…	5
1	Harga gak II jauh beda dgn apps lain. Kelengk…	2
2	Respon lama., mau booking hotel sudah di TF la…	1
3	Sangal mengecewakan saya sudah download aplika…	1
4	Setelah update. login sulit. Sudah fix tiket…	3
….	….	…
495	Reaktif saat rapid test, mengajukan refund lup…	1
496	saya udah mengajukan pembatalan dan refund tik...	1
497	Dari abis update malah ngk bisa proses pembaya…	1
49«	Kecewa. free rapid antigen yg sy dpt dr liket…	1
499	Sangat mengecewakan. saya isi data penumpang b…	1

500 rows x 4 columns

Setelah data ulasan dan score rating berhasil diperoleh dan disaring, langkah selanjutnya adalah menyimpan data tersebut dalam format CSV. Data ini akan digunakan sebagai dataset dalam penelitian seperti pada Tabel 3.

Tabel 3 Data CSV

No	Ulasan	Score
1	Gak bisa pilih mau duduk di kursi yang lebih mahal, langsung diarahkan ke pembayaran.	1
2	Tiket yang telah dipesan mohon untuk bisa diganti metode pembayarannya.	3
3	Reaktif saat rapid test, mengajukan refund lupa lampirkan bukti, mohon bantuannya.	1
…	..	..
499	Dari abis update malah nggak bisa proses pembayaran. Sudah coba ulang tetap gagal.	1

500 rows x 4 columns

Dataset yang telah disimpan kemudian dipanggil kembali dengan memberikan bobot atau nilai data dari tiap score yang diperoleh dari ulasan pengguna [14]. Skor rating tersebut kemudian dikategorikan menjadi dua bagian yaitu positif, dan negatif. Skor Rating 1-2 dikategorikan "negatif", dan Skor Rating 3-5 dikategorikan Positif.

Tabel 4 Pembobotan Dataset

No	Ulasan	Score	Hasil
1	Gak bisa pilih mau duduk di kursi yang lebih mahal, langsung diarahkan ke pembayaran.	1	Negatif
2	Tiket yang telah dipesan mohon untuk bisa diganti metode pembayarannya.	3	Positif
3	Reaktif saat rapid test, mengajukan refund lupa lampirkan bukti, mohon bantuannya.	1	Negatif
..	…	.	…
499	Dari abis update malah nggak bisa proses pembayaran. Sudah coba ulang tetap gagal.	1	Negatif

500 rows x 4 columns

Hasil Preprocessing Data

Preprocessing merupakan tahapan penting dalam menganalisis data yang dilakukan sebelum memulai proses analisis. Tujuan dari *preprocessing* adalah untuk mempersiapkan data agar sesuai untuk proses analisis selanjutnya. Dalam tahap *Preprocessing* Data, terdapat beberapa teknik yang digunakan antara lain:

Cleaning Data difungsikan untuk membersihkan data dari nilai yang tidak valid atau tidak lengkap [15]. Adapun hasil dari proses cleaning dapat dilihat pada Tabel 5.

Tabel 5 Cleaning Data

No	Ulasan Asli	Cleaning Result	Hasil
1	Gak bisa pilih mau duduk di kursi yang lebih mahal, langsung diarahkan ke pembayaran.	gak bisa pilih duduk kursi mahal langsung bayar	Negatif
2	Tiket yang telah dipesan mohon untuk bisa diganti metode pembayarannya.	tiket telah dipesan mohon bisa diganti metode pembayaran	Positif
3	Reaktif saat rapid test, mengajukan refund lupa lampirkan bukti, mohon bantuannya.	reaktif rapid test ajukan refund lupa lampirkan bukti mohon bantuan	Negatif
..	..	..	…
5	Dari abis update malah nggak bisa proses	habis update tidak bisa proses	Negatif

	pembayaran. Sudah coba ulang tetap gagal.	pembayaran sudah coba tetap gagal

Case Folding menurut [16], digunakan dalam mengubah semua huruf menjadi huruf kecil untuk menghilangkan perbedaan dalam sebuah tulisan, dapat dilihat pada Tabel 6.

Tabel. 6. Case folding

No	Ulasan	Score	Hasil
1	gak bisa pilih duduk kursi mahal langsung bayar	1	Negatif
2	tiket telah dipesan mohon bisa diganti metode pembayaran	3	Positif
3	reaktif rapid test ajukan refund lupa lampirkan bukti mohon bantuan	1	Negatif
..	…	..	…
499	habis update tidak bisa proses pembayaran sudah coba tetap gagal	1	Negatif
500 rows x 4 columns			

Tokenizing pada penelitian [17], digunakan dalam pemisahan teks menjadi unit-unit terkecil. Data teks dipotong berdasarkan pemisah seperti spasi atau tanda baca, hasil *tokenizing* ditunjukan pada tabel 7.

Tabel. 7. Tokenizing

No	Ulasan	Score	Hasil	Tokenizing
1	gak bisa pilih duduk kursi mahal langsung bayar	1	Negatif	['gak', 'bisa', 'pilih', 'duduk', 'kursi', 'mahal', 'langsung', 'bayar']
2	tiket telah dipesan mohon bisa diganti metode pembayaran	3	Positif	['tiket', 'telah', 'dipesan', 'mohon', 'bisa', 'diganti', 'metode', 'pembayaran']
3	reaktif rapid test ajukan refund lupa lampirkan bukti mohon bantuan	1	Negatif	['reaktif', 'rapid', 'test', 'ajukan', 'refund', 'lupa', 'lampirkan', 'bukti', 'mohon', 'bantuan']
..	…	..	…	…
49 9	habis update tidak bisa proses pembayaran sudah coba tetap gagal	1	Negatif	['habis', 'update', 'tidak', 'bisa', 'proses', 'pembayaran', 'sudah', 'coba', 'tetap', 'gagal']
500 rows x 4 columns				

Stopword pada penelitian [18], digunakan untuk menghilangkan kata yang tidak memiliki makna khusus. Proses ini dilakukan untuk membersihkan teks untuk meningkatkan kualitas dataset, dan hasilnya dapat dilihat pada tabel 8.

Tabe. 8. Stopwords

No	Tokenizing	Stopword Removal
1	['gak', 'bisa', 'pilih', 'duduk', 'kursi', 'mahal', 'langsung', 'bayar']	['pilih', 'duduk', 'kursi', 'mahal', 'bayar']
2	['tiket', 'telah', 'dipesan', 'mohon', 'bisa', 'diganti', 'metode', 'pembayaran']	['tiket', 'dipesan', 'diganti', 3'metode', 'p…embayaran']
	['reaktif', 'rapid', 'test', 'ajukan', 'refund', 'lupa', 'lampirkan', 'bukti', 'mohon', 'bantuan']	['reaktif', 'rapid', 'test', 'ajukan', 'refund', 'lupa', 'lampirkan', 'bukti', 'bantuan']
…	…	…
499	['habis', 'update', 'tidak', 'bisa', 'proses', 'pembayaran', 'sudah', 'coba', 'tetap', 'gagal']	['update', 'proses', 'pembayaran', 'coba', 'gagal']
500 rows x 4 columns		

Steming berfungsi untuk mengubah kata-kata dalam teks menjadi bentuk dasar atau akar kata. kata-kata dalam teks akan dipangkas atau dipotong sehingga hanya menyisakan bentuk dasar atau akar kata, dan kemudian diubah dalam bentuk string seperti pada tabel 9.

Tabel. 9. Stemming

No	Tokenizing	Stopword Removal	Stemming
1	['gak', 'bisa', 'pilih', 'duduk', 'kursi',	['pilih', 'duduk', 'kursi',	['pilih', 'duduk', 'kursi',

	'mahal', 'langsung', 'bayar']	'mahal', 'bayar']	'mahal', 'bayar']
2	['tiket', 'telah', 'dipesan', 'mohon', 'bisa', 'diganti', 'metode', 'pembayaran']	['tiket', 'dipesan', 'diganti', 'metode', 'pembayaran']	['tiket', 'pesan', 'ganti', 'metode', 'bayar']
3	['reaktif', 'rapid', 'test', 'ajukan', 'refund', 'lupa', 'lampirkan', 'bukti', 'mohon', 'bantuan']	['reaktif', 'rapid', 'test', 'ajukan', 'refund', 'lupa', 'lampirkan', 'bukti', 'bantuan']	['reaktif', 'rapid', 'test', 'aju', 'refund', 'lupa', 'lampir', 'bukti', 'bantu']
…	..	…	..
499	['habis', 'update', 'tidak', 'bisa', 'proses', 'pembayaran', 'sudah', 'coba', 'tetap', 'gagal']	['update', 'proses', 'pembayaran', 'coba', 'gagal']	['update', 'proses', 'bayar', 'coba', 'gagal']
500 rows x 4 columns ,			

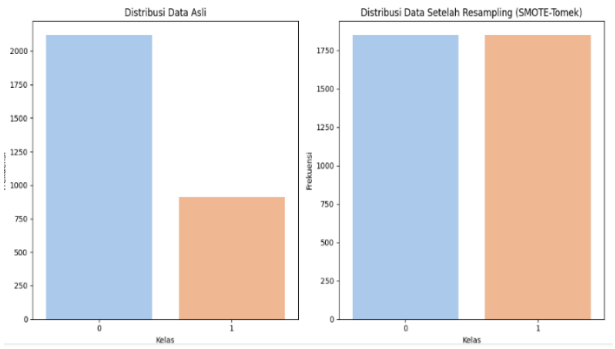
Dapat dilihat pada tabel. 10 dibawah merupakan hasil yang diperoleh output TF yang merupakan frekuensi kemunculan kata pada dokumen. Jika nilainya 0 maka kemunculan kata pada dokumen tidak ada namun jika bernilai 1 maka kemunculan kata pada dokumen tersebut terdapat 1. Hasil TF-IDF ditujukan pada tabel 10.

Tabel. 10. Pembobotan TF-IDF

No	**Term**	**TFA**	**TFB**	**TFC**	**TFD**	**DF**	**IDF (log10(4/DF))**
1	pilih	1	0	0	0	1	0.602
2	dudu k	1	0	0	0	1	0.602
3	kursi	1	0	0	0	1	0.602
4	mahal	1	0	0	0	1	0.602
5	bayar	1	1	0	1	3	0.125
6	tiket	0	1	0	0	1	0.602
7	pesan	0	1	0	0	1	0.602
8	ganti	0	1	0	0	1	0.602
9	metod e	0	1	0	0	1	0.602
10	reaktif	0	0	1	0	1	0.602
11	rapid	0	0	1	0	1	0.602
12	test	0	0	1	0	1	0.602
13	aju	0	0	1	0	1	0.602
14	refun d	0	0	1	0	1	0.602
15	lupa	0	0	1	0	1	0.602
16	lampi r	0	0	1	0	1	0.602
17	bukti	0	0	1	0	1	0.602
18	bantu	0	0	1	0	1	0.602
19	updat e	0	0	0	1	1	0.602
20	proses	0	0	0	1	1	0.602
21	coba	0	0	0	1	1	0.602
22	gagal	0	0	0	1	1	0.602

Pembobotan TF-IDF dilakukan dengan menggunakan sampel untuk perhitungan TF-IDF secara manual. Nilai TF (*Term Frequency*) dapat didapatkan dari tabel 10 di atas yang mencatat frekuensi kemunculan kata-kata dalam setiap dokumen. Sedangkan perhitungan IDF (*Inverse Document Frequency)* dilakukan dengan membagi total jumlah dokumen dengan DF (*Document Frequency*). DF merupakan hasil dari total TF pada masing-masing *Term*.

Setelah menerapkan metode *SMOTE-Tomek Links* pada data, didapatkan hasil yang menghasilkan perbaikan pada masalah ketidakseimbangan kelas dalam dataset. Metode ini telah menghasilkan dataset yang memiliki jumlah data yang seimbang antara kelas mayoritas dan kelas minoritas. Dengan penerapan *SMOTE-Tomek Links*, jumlah data pada kelas minoritas telah ditingkatkan melalui proses oversampling menggunakan SMOTE. Sampel sintetis baru telah dibuat di antara data minoritas yang ada, sehingga meningkatkan representasi kelas minoritas. Selain itu, metode Tomek Links juga telah digunakan untuk menghapus pasangan sampel ambiguitas antara kelas minoritas dan mayoritas. Hal ini membantu meningkatkan pemisahan antara kedua kelas dan mengurangi overlap antara kelas-kelas yang berdekatan dalam ruang fitur. Dengan menggabungkan SMOTE dan Tomek Links, hasil penerapan *SMOTE-Tomek Links* memberikan dataset yang lebih seimbang secara proporsional antara kelas mayoritas dan kelas minoritas. Hal ini dapat memperbaiki kinerja model dalam mempelajari kedua kelas dengan lebih baik, mengurangi bias yang mungkin timbul akibat ketidakseimbangan kelas, dan meningkatkan kemampuan model dalam mengklasifikasikan data yang baru. Berikut merupakan data yang telah di dengan *SMOTETomek Links* pada gambar 2.




```
Total Label Dari Semua data :
{'Negative': 375, 'Positive': 125}
```

Gambar 2 Distribusi Data menggunakan SMOTETomek

Klasifikasi Menggunakan Metode SVM

Data yang telah melakukan penerapan *SMOTETomek links* kemudian diklasifikasikan dalam bentuk SVM. Tujuan dari pengujian ini adalah untuk mengevaluasi kinerja metode SVM dalam mengklasifikasikan data yang telah diubah melalui penerapan *SMOTETomek links*. Pada pengujian ini, dilakukan perbandingan kinerja antara pengklasifikasi SVM yang menggunakan *SMOTETomek Links* dengan yang tidak menggunakan *SMOTETomek Links*. Hal ini bertujuan untuk melihat perbedaan akurasi hasil klasifikasi antara kedua pendekatan tersebut. Berikut ini pada gambar 5 merupakan perbandingan dari visualisasi *confusion matrix* dengan pengklasifikasi SVM yang menggunakan *SMOTETomek Links* dengan yang tidak menggunakan *SMOTETomek Links*.

Tabel. 11. Tabel *Confusion Matrix* Tanpa SMOTE-Tomek Links

Aktual	Terklasifikasi Negatif	Terklasifikasi Positif
Negatif	11	296
Positif	10	289

Perhitungan manual dari hasil dari tabel *confusion matrix* dapat dilihat sebagai berikut:

Accuracy $= \frac{289+11}{289+11+296+10} = \frac{300=}{606}$ 0.49549.5%

Precission $= \frac{289}{289+296} = \frac{289}{285}$ = 0.494 =49.4%

Recall $= \frac{289}{289+10} = \frac{289}{299}$ = 0.967 atau 96.7%

F1-score $= 2 \text{ x } \frac{0.494 \times 0.967}{0.494+0.967}$ = 0.655 atau 65.5%

Tabel. 12. Tabel Confusion Matrix Dengan SMOTE-Tomek Links

Aktual	Terklasifikasi Negatif	Terklasifikasi Positif
Negatif	11	253
Positif	6	259

Perhitungan manual dari hasil dari tabel *confusion matrix* dapat dilihat sebagai berikut :

Accuracy $= \frac{259+11}{259+11+253+6} = \frac{270}{529}$ = 51.0%

Precission $= \frac{259}{259+253} = \frac{259}{512}$ = 50.5%

Recall $= \frac{259}{259+6} = \frac{259}{265}$ = 97.7%

F1-score $= 2 \text{ x } \frac{0,506 \times 0,977}{0,506+0,977}$ = 66.7%

Dari hasil perhitungan *confusion matrix* diketahui bahwa hasil dari Klasifikasi SVM tanpa menggunakan SMOTE-Tomek Links mendapatkan nilai akurasi sebesar 0,49.5 jika dipersenkan akan menjadi 49.5%, sedangkan dengan menggunakan SMOTE-Tomek Links nilai akurasi yang didapatkan sebesar 0,510, dan jika dibulatkan dalam persen mendapatkan hasil sebesar 51%.

Klasifikasi Menggunakan Metode KNN

Tahap berikutnya melibatkan klasifikasi dataset yang telah dibersihkan menggunakan algoritma *K-Nearest Neighbor* (KNN). Pembagian data latih dan data uji dilakukan menggunakan teknik *K-Fold cross validation* dengan jumlah fold *k* sebanyak 3. Teknik ini dipilih untuk memastikan bahwa setiap bagian data diuji secara adil dan hasilnya lebih stabil. Model pembelajaran KNN ini dibangun menggunakan *Python* dengan bantuan *Google Colabs* sebagai platform pengolahan data.

Pengujian algoritma KNN dilakukan dengan menggunakan tiga nilai k untuk tetangga terdekat, yaitu $k = 1m\ k = 2, k = 3$. Penggunaan $k = 3$ dipilih karena memberikan keseimbangan yang baik antara sensitivitas terhadap pola dalam data tanpa terlalu terpengaruh oleh *noise*. Nilai ini cukup kecil untuk menangkap detail penting namun cukup besar untuk menghindari *overfitting*.

Selanjutnya, data yang telah diproses menggunakan metode *SMOTETomek Links* diklasifikasikan menggunakan KNN. Tujuan dari pengujian ini adalah untuk mengevaluasi kinerja metode KNN dalam mengklasifikasikan data yang telah diseimbangkan dengan SMOTETomek. Selain itu, dilakukan perbandingan kinerja antara pengklasifikasi KNN yang menggunakan SMOTETomek Links dengan yang tidak menggunakan SMOTETomek. Pengujian ini bertujuan untuk melihat perbedaan akurasi hasil klasifikasi antara kedua pendekatan tersebut.

Untuk perhitungan manual *Confusion Matrix* dapat dilihat pada tabel 13 dan tabel 14.

Tabel. 13. Tabel Confusion Matrix Tanpa SMOTE-Tomek Links

Aktual	Terklasifikasi Negatif	Terklasifikasi Positif
Negatif	131	176
Positif	131	168

Perhitungan manual dari hasil dari tabel *confusion matrix* dapat dilihat sebagai berikut:

Accuracy $= \frac{168+131}{168+131+176+131} =$
$= 0.49\%$

Precission $= \frac{168}{168+176} = \frac{168}{344}$
$= 48.8\%$

Recall $= \frac{168}{168+131} = \frac{168}{299}$
$= 0.5619$ atau 56.1%

F1-score $= 2 \times \frac{0.4884 \times 0.5619}{0.4884+0.5619} = 0.522$ atau 52.2%

Tabel 14 Tabel Confusion Matrix Dengan SMOTE-Tomek Links

Aktual	Terklasifikasi Negatif	Terklasifikasi Positif
Negatif	143	121
Positif	113	152

Perhitungan manual dari hasil dari tabel *confusion matrix* dapat dilihat sebagai berikut :

Accuracy $= \frac{152+143}{152+143+121+113} = \frac{295}{529}$
$= 0.5577$ atau 55.7%

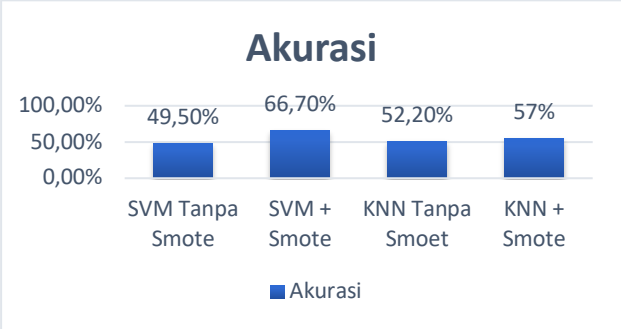
Precission $= \frac{152}{152+121} = \frac{152}{273}$
$= 0.5568$ atau 55.6%

Recall $= \frac{152}{152+113} = \frac{152}{265}$
$= 0.5736$ atau 57.3%

F1-score $= 2 \times \frac{0,5568 \times 0,5736}{0,5568+0,5736} = 0,5651$ atau 56.5%

Confusion Matrix digunakan untuk mengevaluasi kinerja model klasifikasi dengan memberikan gambaran tentang seberapa baik model dapat melakukan prediksi untuk setiap kelas yang ada. Dapat diketahui bahwa hasil dari Klasifikasi KNN tanpa menggunakan SMOTE-Tomek Links mendapatkan nilai akurasi sebesar 0,49.3 jika dipersenkan akan menjadi 49%, sedangkan dengan menggunakan SMOTE-Tomek Links nilai akurasi yang didapatkan sebesar 0,557, dan jika dibulatkan dalam persen mendapatkan hasil sebesar 55%.

Hasil perbandingan akurasi klasifikasi SVM dan KNN dengan penggunaan tanpa *SMOTE-Tomek Links* dan penggunaan *SMOTE-Tomek Links*. Berikut ditunjukan pada gambar 3.



Gambar. 3. Hasil Perbandingan Evaluasi Klasifikasi Metode SVM dan KNN

4. KESIMPULAN

Metode *Support Vector Machine* dan *K-Nearest Neighbor* menunjukkan performa yang baik dalam klasifikasi sentimen, serta efektif dalam mengolah data ulasan dalam jumlah besar dengan tingkat akurasi yang cukup memadai. Hasil pengujian menunjukkan bahwa metode Support Vector Machine tanpa penerapan SMOTE-Tomek Links menghasilkan akurasi sebesar 49,5%, sedangkan setelah menggunakan SMOTE-Tomek Links meningkat menjadi 51%. Sementara itu, metode K-Nearest Neighbor tanpa SMOTE-Tomek Links memperoleh akurasi sebesar 49%, dan meningkat menjadi 55% setelah penerapan teknik penyeimbangan tersebut. Temuan ini menunjukkan bahwa penerapan metode klasifikasi yang tepat, disertai teknik penyeimbangan data, dapat meningkatkan akurasi dalam analisis sentimen ulasan pelanggan.

Namun demikian, penelitian ini memiliki beberapa kelemahan yang perlu diperhatikan. Pertama, akurasi model yang dihasilkan masih tergolong rendah, bahkan setelah penerapan teknik penyeimbangan data, yang menunjukkan perlunya eksplorasi metode klasifikasi lain atau peningkatan pada tahap praproses data. Kedua, data yang digunakan hanya berasal dari satu platform (Tiket.com), sehingga generalisasi terhadap layanan serupa di platform lain masih terbatas. Selain itu, klasifikasi hanya dilakukan dalam tiga label sentimen (positif, negatif, dan netral), tanpa mempertimbangkan nuansa opini yang lebih kompleks seperti sarkasme atau ambiguitas.

DAFTAR PUSTAKA

[1] T. Krisdiyanto, "Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 7, no. 1, hal. 32, 2021, doi: 10.24014/coreit.v7i1.12945.

[2] R. A. Nindy, "Pengaruh Kemudahan Akses dan Diskon Terhadap Keputusan Pembelian

Pada Situs Online Tiket. com (Studi Kasus pada Mahasiswa Aktif S1 Universitas Diponegoro) (Doctoral dissertation, Faculty of Social and Political Sciences).," hal. 1–40, 2018.

[3] M. S. Ummah, "Algoritma Naive Bayes Untuk Prediksi Kepuasan Pelayanan Perekaman E-KTP," *Tri Herdiawan Apanandi*, vol. 11, no. 1, hal. 1–14, 2019.

[4] R. R. Widiaswari, P. Kerja, T. Jawab, dan W. Performance, "Evaluasi Kinerja Pegawai Evaluation of Employee Performance At Barito Kuala District," vol. 5, no. 2, hal. 81–98, 2021, doi: 10.35722/pubbis.

[5] F. Fitriyani, B. Adiwinoto, E. Helmud, A. A. Alkodri, dan A. Anisah, "GDSS Penilaian Kinerja Pada ISB Atma Luhur Menggunakan Metode SAW dan Borda," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, hal. 136–141, 2022, doi: 10.32736/sisfokom.v11i1.1382.

[6] M. N. Muttaqin dan I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," *UNNES J. Math.*, vol. 10, no. 2, hal. 22–27, 2021.

[7] Y. A. Singgalen, "Penerapan Metode CRISP-DM dalam Klasifikasi Data Ulasan Pengunjung Destinasi Danau Toba Menggunakan Algoritma Naïve Bayes Classifier (NBC) dan Decision Tree (DT)," *J. Media Inform. Budidarma*, vol. 7, no. 3, hal. 1551, 2023, doi: 10.30865/mib.v7i3.6461.

[8] D. Surya Sayogo, B. Irawan, dan A. Bahtiar, "Analisis Sentimen Ulasan Instagram Di Google Play Store Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, hal. 3314–3319, 2024, doi: 10.36040/jati.v7i6.8178.

[9] H. M. T. Khushi, T. Masood, A. Jaffar, M. Rashid, dan S. Akram, "Improved Multiclass Brain Tumor Detection via Customized Pretrained EfficientNetB7 Model," *IEEE Access*, vol. 11, no. October, hal. 117210–117230, 2023, doi: 10.1109/ACCESS.2023.3325883.

[10] O. I. Gifari, M. Adha, F. Freddy, dan F. F. S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, no. 1, hal. 36–40, 2022, doi: 10.46229/jifotech.v2i1.330.

[11] H. Said, N. H. Matondang, dan H. N. Irmanda, "Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi," *Techno.Com*, vol. 21, no. 2, hal. 256–267, 2022, doi: 10.33633/tc.v21i2.5901.

[12] A. Cahya Asri dan I. Budiono, "Pelaksanaan Program Indonesia Sehat dengan Pendekatan Keluarga di Puskesmas," *Higeia J. Public Heal. Res. Dev.*, vol. 3, no. 4, hal. 556–567, 2021.

[13] U. Surapati dan A. Y. Zulkarnain, "Implementasi Metode Naïve Bayes Untuk Mendeteksi Hate Speech Pada Twitter," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, hal. 830–837, Sep 2023, doi: 10.31539/intecoms.v6i2.7678.

[14] H. Hamdani dan D. Hartama, "Implementation of Data Analysis Hotel Rating Levels in Bali Using the K-Means Algorithm and Decision Tree," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 10, no. 3, hal. 467–474, 2024, doi: 10.33330/jurteksi.v10i3.2894.

[15] S. Khairunnisa, A. Adiwijaya, dan S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, hal. 406, 2021, doi: 10.30865/mib.v5i2.2835.

[16] A. Sukarno Hatta, "Clustering Pada Data Sentimen Penggunaan Transportasi Online Menggunakan Algoritma Spectral Clustering," *e-Proceeding Eng.*, vol. 8, no. 6, hal. 11945, 2021.

[17] N. Wanaspati, S. Anraeni, dan H. Darwis, "Support Vector Machine untuk Analisis Sentimen Masyarakat Terhadap Penggunaan Antibiotik di Indonesia," *Indones. J. Comput. Sci.*, vol. 12, no. 4, hal. 2196–2206, 2023, doi: 10.33022/ijcs.v12i4.3320.

[18] M. Sukarno Hatta dan F. Azmi, "Clustering Pada Data Sentimen Penggunaan Transportasi Online Menggunakan Algoritma Spectral Clustering," *e-Proceeding Eng.*, vol. 8, no. 6, hal. 11945, 2021.