Journal of Electrical Engineering and Computer (JEECOM)
Vol. 6, No. 2 (2024), DOI: 10.33650/jeecom.v4i2
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Pengaruh Komposisi Split Data Terhadap Performa Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM

Yoga Adi Prasetyo[1], Ema Utami[2], Ainul Yaqin[3]
[1,2,3] Magister Teknik Informatika Universitas Amikom, Yogyakarta, Indonesia

Article Info

Article history:

Diterima 30 Juli 2024
Revisi 5 Agustus 2024
Diterbitkan 4 Oktober 2024

Keywords:

Analisis sentimen
Data Splitting
TF-IDF
SVM
Naïve Bayes

ABSTRAK

Analisis sentimen merupakan bidang yang penting dalam pengolahan bahasa alami dan aplikasi sosial media modern. Penelitian ini menginvestigasi pengaruh dari variasi komposisi split data terhadap performa akurasi model analisis sentimen menggunakan SVM dan Naive Bayes. Metode eksperimen menggunakan variasi dari teknik k-fold cross-validation untuk membandingkan hasil dari berbagai proporsi pembagian data latih dan uji. Hasil eksperimen menunjukkan bahwa komposisi split data memiliki dampak signifikan terhadap performa akurasi kedua algoritma, dengan beberapa proporsi split data menghasilkan hasil yang lebih konsisten dan stabil dibandingkan dengan yang lain. Temuan ini memberikan wawasan yang berharga dalam pengaturan praktis untuk pelatihan model analisis sentimen yang lebih efektif dan andal. Teknik ekstraksi fitur yang digunakan Term Frequency-Inverse Document Frequency (TF-IDF), dengan algoritma klasifikasi Naive Bayes dan Support Vector Machine (SVM). Performa model dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa signifikan model SVM dengan rasio 80:20 mencapai akurasi 76,66% dan F1-score 77 %, dibandingkan metode SVM dan Naïve Bayes dengan rasio lainnya.

Corresponding Author:
Ema Utami,
Magister Teknik Informatika Universitas Amikom, Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia
Email: ema.u@amikom.ac.id

1. PENDAHULUAN

Dalam era digital saat ini, data teks yang dihasilkan dari berbagai platform media sosial, forum, dan blog sangatlah besar. Data ini memberikan peluang yang sangat besar untuk menganalisis opini publik dan sentimen yang terkandung dalam teks-teks tersebut [1]. Analisis sentimen adalah proses untuk mengidentifikasi dan mengklasifikasikan opini atau emosi yang diekspresikan dalam suatu teks untuk mendapatkan berbagai sumber data dari internet dari beragam platform media sosial [2].

Untuk melakukan analisis sentimen, algoritma machine learning seperti Naive Bayes (NB) dan Support Vector Machine (SVM) sering digunakan karena kemampuannya dalam menangani data teks dan memberikan hasil yang akurat. Naive Bayes adalah algoritma probabilistik yang sederhana namun efektif, sedangkan SVM adalah algoritma yang lebih kompleks yang mencari hyperplane optimal untuk memisahkan data dalam ruang fitur berdimensi tinggi [3].

Salah satu langkah krusial dalam pengembangan model machine learning adalah pembagian data (data splitting). Pembagian data yang tepat antara data pelatihan (training data) dan data pengujian (testing data) sangat penting untuk memastikan bahwa model yang dihasilkan memiliki kinerja yang baik dan tidak mengalami overfitting. Overfitting terjadi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga kinerjanya menurun pada data baru yang belum pernah dilihat sebelumnya. Oleh karena itu, pemilihan rasio pembagian data yang tepat dapat membantu dalam menghasilkan model yang lebih generalis dan memiliki kemampuan prediktif yang lebih baik [4].

Penelitian [5] hasil pengujian terhadap 4 algoritma yaitu SVM, Naïve Bayes, Decision Tree, dan Gradient Boosting menunjukkan bahwa algoritma SVM dengan pemisahan data 90:10 merupakan algoritma terbaik untuk klasifikasi sentimen analisis ChatGPT.

Penelitian [6] Penggunaan perbandingan antara persentase data latih dan data uji berpengaruh terhadap peningkatan nilai akurasi yang didapatkan. Nilai akurasi terbaik diperoleh pada rasio perbandingan persentase data latih dan data uji yaitu 90%:10% dengan nilai akurasi hasil klasifikasi menggunakan pembobotan TFRF sebesar 93,8% dan pembobotan TF-IDF sebesar 94,6%. Berdasarkan nilai tertinggi tersebut, dapat disimpulkan bahwa semakin tinggi persentase data latih maka semakin tinggi pula nilai akurasi yang diperoleh.

Penelitian [7] akurasi tertinggi ada pada rasio data latih dan data uji sebesar 90:10, yaitu sebesar 73%. Nilai recall, precision, dan f-measure pada perbandingan 90:10 juga memperoleh hasil paling tinggi, yaitu sebesar 46%, 73%, dan 49%. Hal ini dikarenakan algoritma SVM termasuk dalam metode supervised learning dimana dalam metode tersebut sangat bergantung pada data pelatihan. Semakin banyak data pelatihan maka semakin baik juga akurasi yang dihasilkan.

Penelitian [8] disimpulkan bahwa kinerja metode Naive Bayes pada penelitian ini menggunakan 1000 data review yang diambil dari aplikasi Motorku X di Google Play yang kemudian dibagi menjadi 2 kelas yaitu kelas positif dan negatif. Penelitian dilakukan dengan 3 skenario pembagian data pelatihan dan pengujian yaitu 90%:10%, 80%:20%, dan 70%:30% menghasilkan hasil terbaik pada rasio 90%:10% dengan akurasi 76%. , presisi 76%, dan perolehan 97%.

Penelitian [9] menunjukkan bahwa analisis sentimen dapat dilakukan dengan baik menggunakan Naïve Bayes Classifier dikombinasikan dengan TF-IDF untuk pemilihan fitur. Pada tahap pendahuluan, dengan pemisahan data yang sama sebesar 80:20, skor akurasi sebesar 64,796%, sedangkan skor akurasi pada percobaan utama, ketika jumlah kejadiannya jauh lebih besar, adalah 73,722% sebagai skor terendah. Performanya meningkat dengan akurasi sekitar 8,926%.

Penelitian [10] disimpulkan akurasi tertinggi dicapai oleh Support Vector Machine, dengan hasil akurasi sebesar 93% pada training data dan testing data dengan presentase rasio 70:30. Sementara itu, Naïve Bayes mencapai akurasi sebesar 91,6% pada presentase rasio pengujian yang sama.

Penelitian [11] disimpulkan peningkatan ukuran data training akan meningkatkan kinerja ketiga pengklasifikasi; lebih dari 70% sampel kumpulan data diperlukan dalam fase pelatihan untuk mencapai kinerja yang lebih baik. Penelitian ini berfokus pada analisis dampak dari berbagai rasio pembagian data yang dieksplorasi meliputi 50:50, 60:40, 70:30, 80:20, dan 90:10.

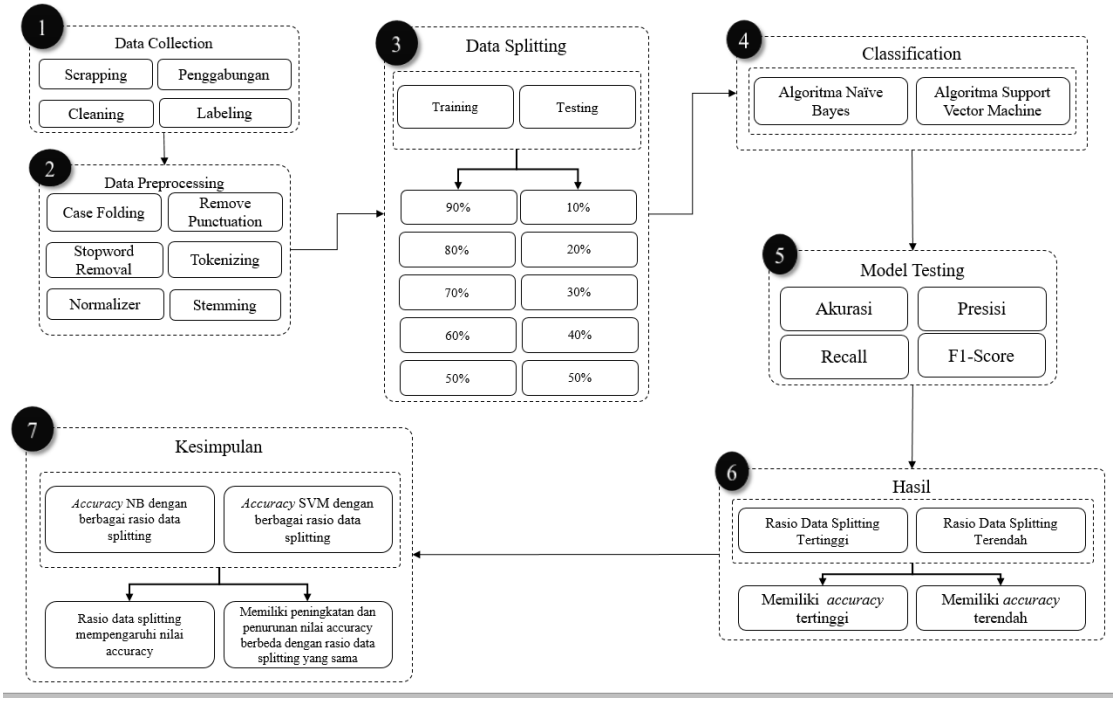
Penelitian [12] menggunakan rasio train – test 50:50, 60:40, 70:30, 80:20, dan 90:10 untuk split. Hasil menunjukkan bahwa menyimpulkan bahwa penggunaan 90:10 memberikan persentase akurasi terbaik dalam tiga pendekatan klasifikasi yang digunakan, diikuti dengan rasio pembagian 70:30 karena memberikan akurasi prediksi yang baik.

Penelitian ini diharapkan dapat memberikan wawasan yang mendalam mengenai bagaimana berbagai rasio pembagian data mempengaruhi performa algoritma Naive Bayes dan SVM dalam analisis sentimen. Dengan demikian, hasil penelitian ini akan memberikan panduan praktis bagi peneliti dan praktisi dalam memilih rasio pembagian data yang paling sesuai untuk diterapkan dalam analisis sentimen, sehingga dapat menghasilkan model yang lebih akurat dan andal.

2. METODE

Ada beberapa langkah yang dilakukan pada penelitian ini, yakni pertama pengumpulan data, preprocessing data dengan menggunakan jenis feature extraction TF IDF, selanjutnya, implementasi algoritma Naïve Bayes

dan *Support Vector Machine* (SVM) dan terakhir adalah *model evaluation* menggunakan *Confusin Matrix*. Langkah-langkah tersebut dipaparkan pada Gambar 1.



Gambar 1. Langkah-langkah penelitian

2.1. Pengumpulan Data

Penelitian ini menggunakan data yang ditambang dari internet, khususnya dari Twitter menggunakan library *auth twitter* google collaboratory. Data yang diperoleh sebanyak 3140 data. Yang kemudian diproses untuk diambil variable penting dari data yakni variabel "full_text" dan "sentimen". Contoh dataset bisa dilihat pada Tabel 1.

Tabel 1. Dataset

TABLE I. DATASET

Full_text
Bromo setelah kebakaran ðŸ¥º, Semoga lekas pulih ðŸŒ» https://t.co/rhfa5dbZbR
Relawan Ganjar pindah ke Prabowo, TAK MAU BONEKA PARTAI Indonesia baekhyun Majalah Bobo Adulting Bromo Chris Evans KDRT Siang Rempang Weverse Selasa Melayu Sotong Bangkok #PRABOWOPRESIDEN2024 #AllinPrabowo #PS08MENANG #BersamaPrabowo #GPKE08 https://t.co/9EFXb6VJ2f
@tanyarlfes Goblok bgt yg nyaranin/ngonsep pake flare, lu pikir lagi di stadion bola, noh sekalian masuk ke kawah bromo.. dapet flare alaminya

2.2. *Preprocessing*

Setelah melewati beberapa tahapan, data yang sudah diperoleh kemudian akan melalui proses *preprocessing*. *Preprocessing* merupakan tahapan awal dalam beberapa tugas pemrosesan teks, termasuk analisis sentimen. *Preprocessing* adalah langkah yang penting dalam analisis sentimen dikarenakan dapat mempengaruhi akurasi hasil secara signifikan [13]. Proses ini dilakukan dengan menggali, mengelola dan menerjemahkan informasi serta menggali hubungan data terstruktur dan tidak menguji data dengan menghilangkan noise dan menyamakan kata serta menambah volume data. Tahapan *preprocessing* meliputi:

- *Case folding* mengubah huruf *uppercase* menjadi huruf *lowercase* [13]
- *Cleaning* pembersihan kata dengan menghilangkan tanda baca [14]
- *Tokenizing* membagi teks menjadi token terpisah untuk menetapkan indeks unik pada setiap kata [15]

- *Stopwords* untuk menghilangkan kata ganti yang tidak memiliki makna [13]
- *Stemming* mendapatkan kata dasar dengan menghilangkan akhiran [16]

2.3. Rasio Data Splitting

Setelah melakukan preprocessing, klasifikasi sentimen dengan feature extaction dengan TF-IDF, proses selanjutnya adalah implementasi algoritma Naïve Bayes dan SVM. Pembagian jumlah data training dan data testing adalah salah satu faktor yang menentukan akurasi. Dataset dibagi menjadi data train dan data test dengan rasio perbandingan yang digunakan pada split data yakni 90:10, 80:20, 70:30,60:40,50:50 [17].

2.4. *Feature Extraction*

Feature extraction menjadi aspek penentu dalam meminimalisir kesalahan informasi [18]. Menurut [19] pendekatan *feature extraction* digunakan untuk mengekstraksi fitur berharga dari kumpulan data berdimensi tinggi.

[20] *Feature Extraction* menjadi proses untuk menganalisis, mengolah, atau mengelola data. Fitur yang diekstraksi harus dalam format tertentu yang dapat langsung menjadi inputan untuk algoritma klasifikasi.

Teknik *feature extraction* diterapkan pada data training dan data testing. Pada data training untuk melatih model yang dipilih dan pada data testing saat klasifikasi dilakukan [21]. Dengan cara yang sama, penggunaan berbagai teknik *feature extraction* terbukti meningkatkan akurasi dari klasifikasi. Beberapa teknik yang sering digunakan adalah *term frequency* (TF), *inverst document frequency* (IDF), TF-IDF, *word2vec* dan *doc2vec* [21].

Pada penelitian ini feature extraction yang digunakan adalah TF-IDF. TF-IDF merupakan metode *feature extraction* yang paling populer dan sering digunakan. TF-IDF dinilai penting, apabila sebuah kata lebih sering muncul dalam sebuah dokumen maka nilai kontribusinya akan semakin besar, namun apabila kata tersebut sering muncul dalam beberapa dokumen maka akan memiliki kontribusi yang lebih kecil [11]. TF-IDF terdiri atas Term Frequency (TF) dan Inverse Document Frequency (IDF) [22]. TF-IDF dapat dirumuskan seperti persamaan berikut :

$$TF\ IDF\ (t_k, d_j) = TF\ (t_k, d_j) * IDF\ (t_k) \quad (1)$$

2.5. Implementasi Algoritma *Naïve Bayes*

Multinomial NBC merupakan model pengembangan dari algoritma bayes yang cocok dalam pengklasifikasian teks atau dokumen. Pada formula Multinomial Naive Bayes Classifier, kelas dokumen tidak hanya ditentukan dengan kata yang muncul tetapi juga jumlah kemunculannya [23].

Kemudian *feature extraction* (TF-IDF) dilatih pada algoritma Naïve Bayes. Proses ini juga akan dilakukan dengan lima skenario secara berulang, berdasarkan hasil klasifikasi dari setiap *rasio data split* yang diimplementasi kemudian performa dari model akan dievaluasi menggunakan *Confusion Matrix*.

2.6. Implementasi Algoritma *Support Vector Machine* (SVM)

Sebelumnya Algoritma *Support Vector Machine* (SVM) merupakan sebuah supervised classifier, diterapkan secara luas untuk menyelesaikan masalah regresi dan klasifikasi. Algoritma ini dirancang sebagai peningkatan untuk *support vector classifier*, yang telah dikenalkan sebagai peningkatan untuk *margin classifier tertinggi*, berurusan dengan data sedarhana dan dapat dipisahkan secara linier memaksimalkan margin antara dua kelas [24].

Kemudian *feature extraction* (TF-IDF) dilatih pada algoritma SVM. Proses ini juga akan dilakukan dengan lima skenario secara berulang, berdasarkan hasil klasifikasi dari setiap *rasio data split* yang diimplementasi kemudian performa dari model akan dievaluasi menggunakan *Confusion Matrix*.

2.7. Evaluasi Algortima

Model divaluasi menggunakan *confusion matrix*. *Confusion matrix* merupakan sebuah alat untuk mengevaluasi kinerja algoritma machine learning yang berisi informasi tentang klasifikasi dan prediksi actual. Ada empat indikator yang diukur di dalamnya: *accuracy, precision, recall* dan *F1-Score* [25]. Pada Gambar 2 adalah skenario dari *Confusion Matrix*.

		Predicted class	
		Class = True	**Class = False**
Actual class	**Class = True**	True positive	False Negative
	Class = False	False Positive	True Negative

Gambar 2. Skenario *Confusion Matrix*

Adapun perhitungan dari 4 indikator adalah sebagai berikut:

Accuracy: $Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)}$ (2)

Recall: $Recall = \frac{TP}{(TP+FN)}$ (3)

Precision: $Precission = \frac{TP}{(TP+FP)}$ (4)

F1-Score :$F1 - Measure = 2 \times \frac{precission \times recall}{precission + recall}$ (5)

Dari proses ini akan ditemukan rasio data split mana yang mampu meningkatkan akurasi dari Naïve Bayes dan SVM pada analisis sentimen.

3. HASIL DAN PEMBAHASAN

Berikut adalah hasil dan pembahasan dari penelitian ini.

3.1. Pengumpulan Data

Dari proses pengumpulan data didapatkan data sebesar 3140 data kemudian diproses untuk diambil variable penting dari data. Yakni variabel "full_text". Data yang diperoleh berbahasa Indonesia

Tabel 2. Hasil Crawling Twitter

TABLE II. HASIL CRAWLING TWITTER

Full_text
Bromo setelah kebakaran ðŸ¥º, Semoga lekas pulih ðŸŒ» https://t.co/rhfa5dbZbR
Relawan Ganjar pindah ke Prabowo, TAK MAU BONEKA PARTAI Indonesia baekhyun Majalah Bobo Adulting Bromo Chris Evans KDRT Siang Rempang Weverse Selasa Melayu Sotong Bangkok #PRABOWOPRESIDEN2024 #AllinPrabowo #PS08MENANG #BersamaPrabowo #GPKE08 https://t.co/9EFXb6VJ2f
@tanyarlfes Goblok bgt yg nyaranin/ngonsep pake flare, lu pikir lagi di stadion bola, noh sekalian masuk ke kawah bromo.. dapet flare alaminya

3.2. *Preprocessing*

Dari keseluruhan tahap pada *preprocessing* yakni, *case folding, cleaning, tokenizing, stopword* dan *stemming*, kami akan berfokus memaparkan hasil dari proses preprocessing pada Tabel 3.

Tabel 3. Hasil *Preprocessing*

TABLE III. CONTOH HASIL *PREPROCESSING*

Text_clean	**sentimen**
bromo setelah kebakaran semoga lekas pulih	Positif
relawan ganjar pindah ke prabowo tak mau dukung capres gagal	Netral
virly virginia indonesia baekhyun majalah bobo adulting bromo…	
lebih gila prewed di bromo sih cuma butuh flare doang	Negatif

3.3. Rasio Data Splitting

Data set dalam penelitian ini merupakan data tanggapan mengenai Kebakaran Savanna Bromo yang diambil dari platform sosial media twitter dari tanggal 1 September 2023 sampai dengan tanggal 31 Desember 2023. Data tanggapan yang diperoleh sebanyak 3149 tanggapan. Penentuan data training dengan perbandingan rasio untuk masing-masing data positif dan negatif bergantung pada rasio perbandingan yang digunakan pada split data yakni 90:10, 80:20, 70:30,60:40,50:50. Hasil akurasi dari setiap jenis berbeda satu sama lain. Berikut adalah hasil akurasi dari data-split.

Tabel 4. Rasio data splitting

TABLE IV. RASIO DATA SPLITTING YANG DIGUNAKAN

Class	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**
Training	1575	1400	1225	1050	875
Testing	176	351	526	701	876

3.4. *Feature Extraction*

Feature extraction menjadi aspek signifikan dari deteksi kesalahan informasi dalam efektivitas algoritma machine learning [18]. Pendekatan *feature extraction* digunakan untuk mengekstraksi fitur berharga dari kumpulan data berdimensi tinggi [19]. Seperti yang dijelaskan pada bagian metode, penelitian ini menggunakan TF-IDF khususnya menggunakan *TfidfVectorizer* pada library sklearn. Cara kerjanya adalah menghitung bobot pada setiap kata dan dapat menyajikan nilai atau skor frekuensi setiap kata dengan tingkat frekuensi kemunculan paling tinggi pada dokumen dalam kasus ini sentimen twitter.

Berikut merupakan hasil kemunculan kata paling tinggi dari proses feature extraction menggunakan TfidfVectorizer yang digunakan.

Tabel 5. Kata dengan nilai TF-IDF tertinggi

TABLE V. HASIL KATA DENGAN NILAI TF-IDF TERTINGGI

kata	**nilai TF-IDF**
bromo	61.4681
kebakaran	52.6093
flare	42.4555
prewedding	41.5769

3.4. Implementasi Algoritma Naïve Bayes dan Support Vector Machine (SVM)

Data set dalam penelitian ini merupakan data tanggapan mengenai Kebakaran Savanna Bromo yang diambil dari platform sosial media twitter dari tanggal 1 September 2023 sampai dengan tanggal 31 Desember 2023. Data tanggapan yang diperoleh sebanyak 3149 tanggapan. Penentuan data training dengan perbandingan rasio untuk masing-masing data positif dan negatif sebanyak 1102 data tanggapan positif dan 247 data tanggapan nengatif. Sehingga total data tanggapan yang digunakan sebanyak 1349 tanggapan. Sedangkan untuk data netral tidak digunakan karena tidak memberi informasi yang penting. Pada penelitian ini seluruh data dibagi kedalam 5 kombinasi perbandingan rasio data training dan data testing seperti pada tabel dibawah ini.

Tabel 6. Nilai Akurasi Data Splitting

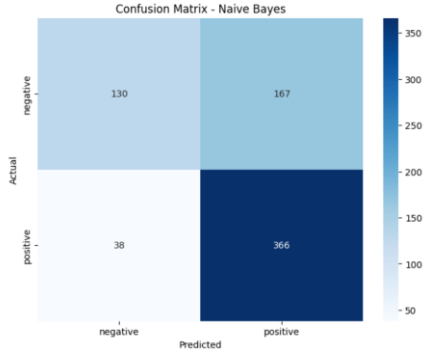
TABLE VI. NILAI ACCURACY BERBAGAI RASIO DATA SPLITTING

Class	**90:10**	**80:20**	**70:30**	**60:40**	**50:50**
Naïve Bayes	70,45 %	71,22 %	70,34 %	70,75 %	68,94 %
SVM	75,00 %	**76,63 %**	74,90 %	73,18 %	73,40 %

3.5. Evaluasi Algoritma

Evaluasi algoritma merupakan tahap terakhir setelah proses implementasi model algortima Naïve Bayes dan SVM. Tahap ini dilakukan untuk mengukur performa terhadap model Naïve Bayes dan SVM dengan

menggunakan confusion matrix seperti yang dipaparkan pada bagian metode.. Berikut adalah hasil confusion matrix dari setiap skenario yang dilakukan.



Gambar 3. Confussion Matrix Naïve Bayes

Gambar 4. Confussion Matrix SVM

Pada Tabel 7 akan dipaparkan nilai Accuracy, Precision, Recall dan F1-Score dari lima skenario yang dilakukan pada penelitian ini.

Tabel 7. Accuracy, Precision, Recall dan F1-Score

TABLE VII. NILAI ACCURACY, PRECISION, RECALL DAN F1-SCORE

Skenario	**Accuracy**	**Precision**	**Recall**	**F1-Score**
NB 90:10	70,45 %	73,00 %	70,00 %	69,00 %
SVM 90:10	75,00 %	75,00 %	75,00 %	75,00 %
NB 80:20	71,22 %	74,00 %	71,00 %	70,00 %
SVM 80:20	**76,63 %**	**77,00 %**	**77,00 %**	**77,00 %**
NB 70:30	70,34 %	73,00 %	70,00 %	68,00 %
SVM 70:30	74,90 %	75,00 %	75,00 %	75,00 %
NB 60:40	70,75 %	72,00 %	71,00 %	69,00 %
SVM 60:40	73,18 %	73,00 %	73,00 %	73,00 %
NB 50:50	68,94 %	71,00 %	69,00 %	66,00 %
SVM 50:50	73,40 %	73,00 %	73,00 %	73,00 %

Dari semua skenario yang dilakukan, skenario antara SVM rasio 80:20 mendapatkan nilai Accuracy, Precision, Recall dan F1-Score paling tinggi dari skenario lainnya.

4. KESIMPULAN

Hasil penelitian tentang analisis sentimen yang menggunakan data yang didapatkan dari hasil crawling data Twitter dengan topik "kebakaran savana Bromo" dengan data sebanyak 3140 baris data. Berdasarkan tujuan penelitian ini, yakni mengeksplorasi hasil klasifikasi dengan rasio pembagian data yang berbeda untuk kelas positif, negatif, dan netral, dapat disimpulkan bahwa rasio pembagian data mempengaruhi performa model algoritma yang digunakan, dalam hal ini Naive Bayes dan SVM. SVM rasio 80:20 mendapatkan nilai Accuracy, Precision, Recall, dan F1-Score tertinggi dengan nilai Accuracy sebesar 76,63%, Precision sebesar 77%, Recall sebesar 77%, dan F1-Score sebesar 77%. Saran penelitian selanjutnya, dilakukan eksperimen menggunakan gabungan antara jenis feature extraction yang berbeda pada klasifikasi sentimen serta menggunakan dataset dan algoritma machine learning yang berbeda untuk menghasilkan eksperimen yang lebih kompleks dan teruji.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pembimbing, yakni Ibu Prof. Dr. Ema Utami, S.Si., M.Kom. dan Bapak Ainul Yaqin, M.Kom. Karena sudah membimbing penulis dalam menyelesaikan penelitian ini dengan memberikan saran, pendapat serta memotivasi penulis selama penelitian ini dikerjakan.

REFERENSI

[1] Y. Rohmiyati, "Analisis Penyebaran Informasi Pada Sosial Media," *Anuva*, vol. 2, no. 1, p. 29, 2018, doi: 10.14710/anuva.2.1.29-42.

[2] M. T. Nitamia and H. Februariyanti, "Analisis Sentimen Ulasan Ekpedisi J&T Expres Menggunakan Algoritma Naive Bayes," *J. Manaj. Inform. Sist. Inf.*, vol. 5, no. 1, pp. 20–29, 2022.

[3] U. Khaira, R. Aryani, and R. W. Hardian, "Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Kebijakan Kemdikbudristek Mengenai Kuota Internet Selama Covid-19," *J. Process.*, vol. 18, no. 2, pp. 272–285, 2023, doi: 10.33998/processor.2023.18.2.897.

[4] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *J. Sains dan Inform.*, vol. 9, no. April, pp. 19–28, 2023, doi: 10.34128/jsi.v9i1.622.

[5] S. Rabbani, D. Safitri, F. Try Puspa Siregar, R. Rahmaddeni, and L. Efrizoni, "Evaluation of Support Vector Machine, Naive Bayes, Decision Tree, and Gradient Boosting Algorithms for Sentiment Analysis on ChatGPT Twitter Dataset," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 1, p. 11, 2023, doi: 10.24014/ijaidm.v7i1.24662.

[6] D. N. N. Husnina, D. E. Ratnawati, and B. Rahayudi, "Analisis Sentimen Pengguna Aplikasi RedBus berdasarkan Ulasan di Google Play Store menggunakan Metode Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, pp. 737–743, 2023, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/12297%0Ahttps://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/12297/5593.

[7] C. B. Vista, O. M. Sihono, and A. T. Firdausi, "Analisis Sentimen Kebijakan Pembelajaran Tatap Muka Selama Pandemi Covid-19 Menggunakan Metode Support Vector Machine," *J. Inform. Polinema*, vol. 9, no. 3, pp. 259–264, 2023, doi: 10.33795/jip.v9i3.1273.

[8] A. Mustolih, P. Arsi, and P. Subarkah, "Sentiment Analysis Motorku X Using Applications Naive Bayes Classifier Method," *Indones. J. Artif. Intell. Data Min.*, vol. 6, no. 2, p. 231, 2023, doi: 10.24014/ijaidm.v6i2.24864.

[9] J. Li, M. Ayu, S. T. Albarhami, and K. Kyritsis, *Advances in Sentiment Analysis*, no. January. 2024.

[10] G. paksi Permana, D. A. Nugraha, and H. Santoso, "JOINTECS Perbandingan Performa SVM dan Naïve Bayes Pada Analisis Sentimen," vol. 7, no. 1, pp. 4–6, 2024.

[11] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 331–339, 2024, doi: 10.14569/IJACSA.2024.0150235.

[12] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th Int. Mardin Artuklu Sci. Res. Conf.*, no. February, pp. 496–504, 2022, [Online]. Available: https://www.researchgate.net/publication/358284895_IDEAL_DATASET_SPLITTING_RATIOS_IN_MACHINE_LEARNING_ALGORITHMS_GENERAL_CONCERNS_FOR_DATA_SCIENTISTS_AND_DATA_ANALYSTS.

[13] F. Resyanto, Y. Sibaroni, and A. Romadhony, "Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case:iPhone Tweets," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 2–6, 2019, doi: 10.1109/ICIC47613.2019.8985943.

[14] Y. Handayani, A. R. Hakim, and Muljono, "Sentiment analysis of Bank BNI user comments using the support vector machine method," *Proc. - 2020 Int. Semin. Appl. Technol. Inf. Commun. IT Challenges Sustain. Scalability, Secur. Age Digit. Disruption, iSemantic 2020*, pp. 202–207, 2020, doi: 10.1109/iSemantic50169.2020.9234230.

[15] B. AlBadani, R. Shi, and J. Dong, "A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM," *Appl. Syst. Innov.*, vol. 5, no. 1, 2022, doi: 10.3390/asi5010013.

[16] N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," *SN Comput. Sci.*, vol. 3, no. 1, pp. 1–20, 2022, doi: 10.1007/s42979-021-00958-1.

[17] Baiq Nurul Azmi, Arief Hermawan, and Donny Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, 2023, doi: 10.35746/jtim.v4i4.298.

[18] Y. Barve, J. R. Saini, K. Pal, and K. Kotecha, "A Novel Evolving Sentimental Bag-of-Words Approach for Feature Extraction to Detect Misinformation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 266–275, 2022, doi: 10.14569/IJACSA.2022.0130431.

[19] P. Verma, A. Dumka, A. Bhardwaj, and A. Ashok, "Product Review-Based Customer Sentiment Analysis Using an Ensemble of mRMR and Forest Optimization Algorithm (FOA)," *Int. J. Appl. Metaheuristic Comput.*, vol. 13, no. 1, pp. 1–21, 2022, doi: 10.4018/ijamc.2022010107.

[20] M. B. Ressan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 375–383, 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[21] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Tweets classification on the base of sentiments for US airline

companies," *Entropy*, vol. 21, no. 11, pp. 1–22, 2019, doi: 10.3390/e21111078.
[22] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 422, 2021, doi: 10.30865/mib.v5i2.2845.
[23] A. H. Setianingrum, D. H. Kalokasari, and I. M. Shofi, "Implementasi Algoritma Multinomial Naive Bayes Classifier," *J. Tek. Inform.*, vol. 10, no. 2, pp. 109–118, 2018, doi: 10.15408/jti.v10i2.6822.
[24] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.
[25] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 2, p. 112, 2020, doi: 10.20473/jisebi.6.2.112-122.