JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 1, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i1.5685

ANALISIS SENTIMEN PUBLIK TERHADAP KARAPAN SAPI DI TWITTER MENGGUNAKAN KOMBINASI METODE K-MEANS DAN SUPPORT VECTOR MACHINE

Zalzabila Rani[1], Bain Khusnul Khotimah [2*]

[1,2]Program Studi Teknik Informatika, Universitas Trunojoyo Madura

Received: 13 Desember 2024
Accepted: 14 Januari 2025
Published: 20 Januari 2025

Keywords:
Analisis Sentimen, Twitter, Karapan Sapi, K-Means, Support Vector Machine.

Corespondent Email:
rzalzabila@gmail.com

Abstrak. Perkembangan teknologi dan internet telah mengubah cara komunikasi masyarakat, dengan media sosial menjadi platform utama. Twitter, yang memiliki 18,45 juta pengguna di Indonesia pada 2022, digunakan dalam penelitian ini untuk menganalisis sentimen publik terkait Karapan Sapi, kompetisi balap sapi tradisional di Madura. Penelitian ini bertujuan untuk mengidentifikasi sentimen publik terhadap Karapan Sapi, mengevaluasi efektivitas kombinasi metode K-Means dan Support Vector Machine (SVM), serta menerapkan Synthetic Minority Oversampling Technique (SMOTE) untuk menangani ketidakseimbangan data. Sebanyak 647 ulasan Twitter berbahasa Indonesia dikumpulkan melalui crawling berbasis Python dan diproses menggunakan text preprocessing. Metode K-Means mengelompokkan ulasan menjadi tiga cluster: aspek budaya Karapan Sapi, olahraga tradisional, dan keterkaitan dengan pihak militer atau pemerintah. SMOTE berhasil menyeimbangkan data sentimen, meningkatkan performa model pembelajaran mesin pada kelas minoritas. Model SVM dengan parameter optimal (kernel linear, C=1.0, gamma=1.0) menghasilkan akurasi 92%, meski masih menunjukkan ketidakseimbangan performa antar kelas. Penelitian ini membuktikan efektivitas kombinasi K-Means dan SVM, serta pentingnya SMOTE dalam analisis sentimen berbasis Twitter, khususnya untuk budaya lokal seperti Karapan Sapi.

Abstract. *The development of technology and the internet has changed the way people communicate, with social media being the main platform. Twitter, which had 18.45 million users in Indonesia in 2022, is used in this study to analyze public sentiment related to Karapan Sapi, a traditional cattle racing competition in Madura. This study aims to identify public sentiment towards Karapan Sapi, evaluate the effectiveness of the combination of K-Means and Support Vector Machine (SVM) methods, and apply Synthetic Minority Oversampling Technique (SMOTE) to handle data imbalance. A total of 647 Indonesian Twitter reviews were collected through Python-based crawling and processed using text preprocessing. The K-Means method grouped the reviews into three clusters: cultural aspects of Karapan Sapi, traditional sports, and links to the military or government. SMOTE successfully balanced the sentiment data, improving machine learning model performance on minority classes. The SVM model with optimized parameters (linear kernel, C=1.0, gamma=1.0) produced 92% accuracy, although it still showed imbalance in performance between classes. This research proves the effectiveness of the combination of K-Means and SVM, and the importance of SMOTE in Twitter-based sentiment analysis, especially for local cultures such as Karapan Sapi.*

1. PENDAHULUAN

Twitter merupakan salah satu platform media terkemuka yang sangat disukai dan banyak digunakan untuk mengekspreikan pikiran. Didirikan oleh *Jack Dorsey*, Twitter memberdayakan penggunanya untuk membuat dan berbagi pesan di dalam ranah tweet dan retweet [1].

Istilah "Karapan Sapi" mengacu pada jenis khusus acara balap sapi. Ras ini berasal dari Pulau Madura dan menunjukkan sepasang sapi diikat ke gerobak dengan seorang joki. Penggunaan taji untuk mendorong sapi untuk berakselerasi ke kecepatan maksimal mereka menuju tujuan yang telah ditetapkan membantu proses pengelolaan sapi ini. Biasanya, garis finish berada pada jarak 100 meter dan berlangsung sekitar satu menit. Istilah "Karapan Sapi" berasal dari padanan bahasa Inggris yang disebut "Perlombaan Kuda", yang dianggap unik untuk wilayah Madura, Indonesia. Ini mirip dengan Drag Race, di mana dua pasangan sapi bersaing untuk kecepatan sampai mencapau garis finis [2].

Penelitian sebelumnya telah menggunakan metode SVM untuk melakukan pelabelan data. Maulana Baihaqi et al., [3] melakukan penelitian yang menunjukkan bahwa Twitter dapat digunakan sebagai sumber untuk membuat korpus kalimat yang mengandung elemen SARA. Setelah validasi oleh pakar bahasa, proses pelabelan dan klasifikasi menggunakan algoritme K-Means keduanya sangat akurat. Akurasi awal metode validasi lima kali lipat dan sepuluh kali lipat masing-masing 64,18% dan 63,68%, tetapi setelah data diperbaiki, meningkat menjadi 70,15% dan 71,14% [3]. Hermansyah & Sarno [4] penelitian ini menganalisis produk dan evaluasi pelayanan di PT Telekomunikasi Indonesia menggunakan TextBlob untuk pelabelan dan algoritme Naïve Bayes dan K-NN untuk klasifikasi. Hasilnya menunjukkan bahwa TextBlob mencapai akurasi sebesar 54,67%, Naïve Bayes 69,44%, dan K-NN mencapai akurasi tertinggi sebesar 75% [4]. Menurut beberapa penelitian di atas, algoritme K-Means yang dikombinasikan dengan klasifikasi SVM memberikan hasil pelabelan terbaik, dengan akurasi yang meningkat secara signifikan setelah validasi pakar bahasa, mencapai 71,14%. Ini menunjukkan bahwa algoritme ini merupakan kombinasi metode yang sangat baik untuk analisis sentimen dan pelabelan data [5].

Di bawah ini adalah contoh penelitian terkait yang membahas penggunaan berbagai teknik klasifikasi: Pamungkas & Kharisudin [6] melakukan penelitian yang menganalisis sentimen masyarakat Indonesia terhadap pandemi COVID-19 di Twitter dengan menggunakan tiga algoritma: Naive Bayes, SVM, dan KNN. Algoritma-algoritma ini dibandingkan untuk menentukan algoritma mana yang paling efektif dalam mengklasifikasikan data respons. Algoritma SVM dengan kernel linier memiliki akurasi rata-rata tertinggi sebesar 90,01 %, menurut evaluasi yang dilakukan menggunakan teknik validasi cross-fold 10-fold. Di sisi lain, Naive Bayes dengan smoothing Laplace 1 memiliki akurasi rata-rata 79,20 %, dan KNN dengan nilai K 20 dan kernel optimal memiliki akurasi rata-rata 62,10% [6]. Teknik Komputer AMIK BSI et al., [7] penelitian ini bertujuan untuk mengklasifikasikan tweet calon gubernur Jawa Barat periode 2018-2023 ke dalam kategori positif dan negatif menggunakan algoritma *Naive Bayes* dan SVM, dengan seleksi fitur yang dilakukan menggunakan Genetic Algorithm. Pada dataset yang tidak seimbang, hasil evaluasi menunjukkan bahwa SVM menghasilkan akurasi rata-rata 92,61% dengan AUC 0,950, sementara Naive Bayes menghasilkan akurasi rata-rata 93,29% dengan AUC 0,525. Hasil ini menunjukkan bahwa SVM dapat digunakan untuk mendeteksi kategori positif dan negatif pada tweet dengan akurasi tinggi, khususnya untuk dataset berbahasa Indonesia [7].

Penelitian ini bertujuan untuk menggunakan kombinasi metode K-Means dan SVM dalam menganalisis sentimen publik di Twitter terhadap karapan sapi. Opini masyarakat ini diambil dari komentar dan pendapat yang disampaikan melalui jejaring sosial media yaitu Twitter yang mana tujuannya untuk mengetahui penerapan metode K-Means dan SVM.

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen adalah metode atau pendekatan yang digunakan untuk memahami bagaimana suatu perasaan digambarkan dalam teks [8]. Karena terdapat sekumpulan dokumen

teks yang berisi informasi atau gagasan tentang sesuatu, maka tujuan penambangan adalah mengekstrak ciri-ciri dan aspek dari apa yang tertulis dalam setiap dokumen, kemudian memutuskan apakah itu komentar positif, negatif, atau netral [9] Jadi, sentiment analysis berfokus di pengolahan opini yang mengandung polaritas, yaitu mempunyai nilai sentimen positif, negatif, atau netral [10].

2.2 Clustering K-Means

James B. Mac Queen memperkenalkan metode K-Means untuk pertama kalinya pada tahun 1967 dalam *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Metode ini dianggap sebagai salah satu teknik pembelajaran tanpa pengawasan yang paling sederhana dan efektif untuk menyelesaikan masalah clustering. [11].

Masukan : kelompok dokumen D yang terdiri dari d1 , d2 , d3 , ... dn
Jumlah *cluster* (k) yang akan dibentuk;
Keluaran : k *cluster*;
Proses :

a. Pilih secara acak k dokumen untuk digunakan *centroid* (titik pusat *cluster*) awal;
b. Gunakan persamaan *cosine similarity* untuk menghitung jarak setiap dokumen ke *centroid*, kemudian kelompokkan dokumen-dokumen ke dalam *cluster* berdasarkan jarak terdekat dengan *centroid;*
c. Tentukan *centroid* baru dengan menghitung rata-rata data dari *cluster* tersebut;
d. Ulangi langkah 2 jika posisi *centroid* baru berbeda dengan *centroid* sebelumnya;

Dengan menggunakan alat *RapidMiner*, model yang disarankan dapat digunakan untuk memformulasikan proses *clustering* menggunakan algoritma K-Means. Proses ini mencakup dokumen proses dari berkas, *clustering*, *clustering* peta, dan kinerja. Dari model ini, kita akan mendapatkan nilai akurasi, *presisi*, dan *recall* dari algoritma K-Means.

2.3 SMOTE

Salah satu metode oversampling atau penambahan data minoritas dengan menambahkan sintesis bertujuan untuk menyeimbangkan data minoritas dengan data mayoritas. Teknik ini dikenal sebagai metode smote, yang merupakan singkatan dari teknik oversampling sintetis minoritas. Jika data tidak seimbang antara data minoritas dan mayoritas, proses klasifikasi akan gagal dengan lebih baik. Metode ini menggabungkan informasi baru dari kelas minoritas untuk meningkatkan jumlah kasus dalam kumpulan data yang tidak seimbang. Ketidakseimbangan kelas, atau ketidakseimbangan kelas, dalam klasifikasi pembelajaran mesin dapat diatasi dengan menggunakan metode Smote. Karena SMOTE tidak selalu dapat menjamin peningkatan akurasi model, perlu dilakukan eksperimen dengan berbagai persentase, set fitur, dan jumlah tetangga terdekat yang berbeda untuk mengevaluasi dampak penambahan kasus terhadap kinerja model. Selain itu, SMOTE juga dapat digunakan pada praproses untuk meningkatkan akurasi model yang sedang digunakan. Metode SMOTE dapat digunakan bersama.

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (4)$$

Keterangan :
x_syn = Data sinteis yang akan dibuat
x_i = Data yang akan direplikasi
x_knn = Data tetangga terdekat
δ = Nilai random antara 0 dan 1

2.4 Support Vector Machine (SVM)

Untuk masalah regresi dan klasifikasi, teknik prediksi yang digunakan adalah Support Vector Machine (SVM). Mencari hyperplane optimal dengan memaksimalkan margin adalah cara kerja SVM. Ini bertujuan untuk memisahkan dua kelas dari ruang input. Hyperplane berbentuk garis dalam ruang dua dimensi, tetapi bisa berupa bidang datar (flat plane) atau lebih banyak bidang (multiplane) dalam ruang yang lebih besar [6]. Dibandingkan dengan metode klasifikasi lainnya, SVM lebih mudah dipahami dan lebih terstruktur untuk klasifikasi atau mendukung regresi vektor [12].

Dalam melakukan klasifikasi SVM dengan kernel linear terdapat beberapa langkah Berikut ini proses klasifikasi dengan *Support Vector Machine* menggunakan kernel linear yaitu:

1. Inisialisasi α , C, epsilon, lamda dan gamma.
2. masukan data latih berdasarkan kemunculan satu kata kunci dalam kalimat.
3. Hitung dot product untuk setiap data dengan menggunakan fungsi kernel [K]. Formula fungsi kernel linear adalah sebagai berikut :

$$K(x, y) = \sum nxi \times yi \quad (5)$$

4. Hitung matriks dengan formula sebagai berikut :

$$KDij = Yi\,Yj\,(K\,(Xi\,Xj) + \lambda 2) \quad (6)$$

Keterangan:
Dij = Elemen matriks ke-I,j
Yi = Kelas data ke-i
Yj = Kelas data ke-j

5. Hitung nilai eror dengan formula:

$$E_i = \sum_{j=1}^{i} \propto_j D_{ij} \quad (7)$$

Keterangan :
Ei = nilai eror data ke-i
αj = nilai alfa ke-j
Dij = Matriks Hessian

6. Hitung nilai dari delta alpha dengan formula :

$$\alpha_i = \min\{\max[\gamma(1 - E_i) - E_i]\, C - \alpha_i \quad (8)$$

Keterangan:
αi = alfa nilai ke – i
γ = gamma untuk mencari kecepatan
Eij = rata-rata eror
C = untuk menentukan batas nilai alfa

7. Hitung nilai alpha baru dengan formula

$$\alpha_i = \alpha_i + \delta\alpha_i \quad (9)$$

Keterangan.:
αi = alfa nilai ke – i
$\delta\alpha i$ = delta alfa nilai ke – i

8. Hitung nilai bias dengan formula

$$b = -\frac{1}{2}(w.x^{+} + w.x^{x}) \quad (10)$$

Keterangan:
Wi+ adalah bobot dot product dengan alpha terbesar di kelas positif.
Wi- adalah bobot dot product dengan a lpha terbesar di kelas negatif.

9. sesudah nilai α, w dan b diketahui, maka dilanjutkan ketahap pengujian. Untuk melakukan pengujian dilakukan perhitungan *dot product* antara data test dengan semua data train dengan fungsi kernel rbf. Setelah itu pengujian dilakukan dengan fungsi keputusan:

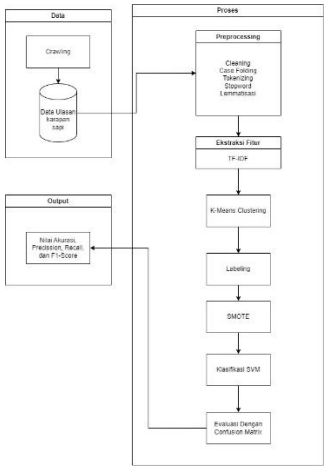
$$f(x) = xw.x + b\ atau\ f(x) = \sum_{j=1}^{m} sign(\propto_j x_i\, k(x, x_i)) + b \quad (11)$$

Keterangan :
αi = alfa nilai ke-i
xi = data nilai dari kelas.ke-i m.= data jumlah dari SV
$K(x, xi)$ =fungsi kernel yang digunakan b = nilai bias

3. METODE PENELITIAN

3.1. Tahapan Penelitian

Pada Gambar 1. Proses analisis sentimen dengan K-Means dan SVM dimulai dengan crawling data ulasan. Setelah preprocessing membersihkan teks, fitur penting diekstraksi menggunakan TF-IDF. Data kemudian dikelompokkan dengan K-Means dan diberi label sentimen. Metode SMOTE digunakan untuk mengatasi ketidakseimbangan data sebelum SVM mengklasifikasikannya. Evaluasi menggunakan matriks konfusi menghasilkan metrik seperti akurasi, presisi, recall, dan skor F1.



Gambar 1 Arsitektur Sistem

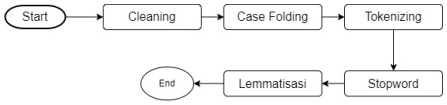
3.1.1. Pengumpulan Data

Data dikumpulkan menggunakan kata kunci "karapan sapi" dari jejaring sosial Twitter dari Maret 2023 hingga Juni 2024.

3.1.2. Preprocessing Data

Prepocessing dilakukan dengan tujuan untuk memudahkan pengolahan data karena

penilaian yang diperoleh selanjutnya akan lebih terstruktur.



Gambar 2 Flowchart Preprocessing

1. *Cleaning* yaitu *p*roses menghilangkan karakter-karakter yang tidak relevan atau mengganggu dalam teks, seperti tanda baca, angka, atau karakter khusus.
2. *Case Folding* pada tahap ini , teks dengan huruf besar diubah menjadi teks huruf kecil secara keseluruhan .
3. *Tokenizing* adalah proses membagi teks menjadi kata-kata khusus, juga dikenal sebagai token.
4. *Stopword Removal* suatu proses penghapusan kata-kata yang umum dan tidak membawa banyak informasi, seperti "dan", "atau", "yang".
5. *Lemmatisasi* adalah proses normalisasi, tetapi fokusnya lebih spesifik, yaitu mengubah kata menjadi bentuk dasar (lemma) dengan memperhatikan konteks gramatikal.

3.1.3. Pembobotan Kata

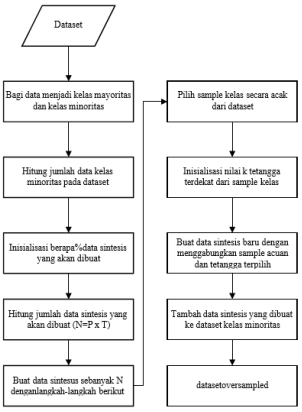
Setelah tahap *text preprocessing*, dataset yang telah diproses akan digunakan untuk ekstraksi fitur metode TF-IDF (Term Frequency-Inverse Document Frequency) yang digunakan. Proses ini mengubah kumpulan kata atau istilah dalam teks menjadi representasi numerik berupa matriks vektor.Ektraksi fitur dilakukan dengan cara memberikan bobot pada setiap kata menghitung jumlah kemunculan kata dalam setiap data.

1. Menghitung frekuensi kemunculan kata (Term Frequency, TF): menghitung jumlah kata yang muncul dalam setiap dokumen.
2. Menghitung nilai IDF (Inverse Document Frequency): Menggunakan rumus IDF untuk menentukan bobot kata berdasarkan keberadaannya di seluruh dokumen dalam dataset.
3. Menghitung bobot TF-IDF: Mengalikan nilai TF dan IDF untuk setiap kata di setiap dokumen.

Hasil akhirnya berupa vektor numerik yang merepresentasikan bobot setiap kata dalam dokumen. Vektor ini dapat digunakan untuk melatih model pembelajaran mesin atau untuk analisis lebih lanjut.

3.1.4. SMOTE

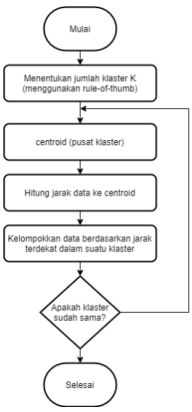
Teknik smote ini digunakan untuk mengatasi ketidak seimbangan data kelas mayoritas dan kelas minoritas. Pada gambar 3 mendeskripsikan lankah-langkah pada Smote. Untuk membuat data sintetis menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*), metode TF-IDF digunakan untuk membagi kata-kata dalam dataset.



Gambar 3 Flowchart SMOTE

3.1.5. K-Means Clustering

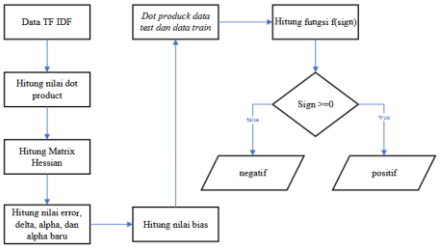
Pada tahap ini, data yang sudah diproses dikelompokkan menggunakan algoritma K-Means untuk mengelompokkan data ke dalam berbagai kelompok berdasarkan karakteristik yang serupa. Tujuan dari tahap ini adalah mengidentifikasi kelompok alami dalam data, seperti sentimen positif atau negatif.



Gambar 4 Flowchart K-Means

3.1.6. Klasifikasi SVM

Dalam penelitian ini, proses klasifikasi menggunakan model *Support Vector Machine* (SVM). Data yang sebelumnya telah diproses dan dihitung bobotnya menggunakan TF-IDF akan dibagi menjadi dua bagian: data *train* dan data *test*.



Gambar 5 Flowchart SVM

3.1.7. Evaluasi

Confusion Matrix merupakan metode perhitungan performa suatu metode klasifikasi dalam memprediksi sentiment positif dan negatif terhadap dataset. Pada *confusion matrix* akan dilakukan tahap evaluasi dengan melakukan perhitungan nilai *accuracy, precision, recall, dan fi-score.*

4. HASIL DAN PEMBAHASAN

Penelitian ini dikembangkan menggunakan *Google Colab* sebagai editor kode, dengan *Python* sebagai bahasa pemrograman. Proses penelitian, mulai dari crawling data hingga evaluasi, memanfaatkan berbagai *lybrary Python*.

4.1 Pengumpulan Data

Dalam penelitian ini, data dikumpulkan melalui teknik crawling data ulasan aplikasi Twiteer. Teknik ini mengumpulkan data mentah sebanyak 647 ulasan dalam format csv.

Gambar 6 Hasil Crawling

Pada gambar 6. menunjukan beberapa fitur awal setelah *crawling* data, hanya data ulasan dengan atribut full_text yang menggunakan bahasa Indonesia yang akan digunaan untuk proses ke tahap selanjutnya.

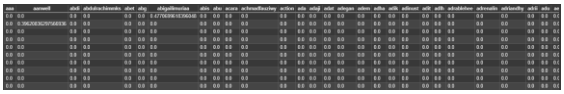
4.2 Text Prepocessing

Hasil *crawling* tersebut dinamakan dataset yang akan melakukan *preprocessing* untuk mengelola teks mentah menjadi lebih terstruktur untuk diproses ketahapan selanjutnya. Tahapan teks *preprocessing* melaui beberapatahapan yang meliputi *tokenizing, clean & case folding, normalization, stopword removal, lemmatization*.

Gambar 7 Hasil Preprocessing

4.3 Pembobotan Kata TF-IDF

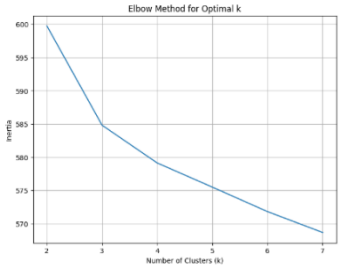
Selanjutnya, metode TF-IDF digunakan untuk membobot kata. Proses ini menghitung berapa banyak kata atau istilah yang muncul dalam setiap dokumen, dan kemudian nilai-nilai ini dianggap sebagai kata bobot.



Gambar 8 Hasil Pembobotan Kata TF-IDF

4.4 Pelabelan *Clustering* K-Means

Pelabelan dengan K-Means adalah metode tanpa pengawasan untuk mengelompokkan data berdasarkan kesamaan fitur tanpa label awal. Dalam pemrosesan teks, teknik ini mengelompokkan dokumen serupa, dengan jumlah cluster ditentukan menggunakan metode Elbow



.

Gambar 9 Hasil Metode Elbow

Berdasarkan Gambar 9, nilai k optimal adalah 3 karena penurunan inertia paling tajam

terjadi antara k=2 dan k=3. Setelah k=3, penurunan menjadi stabil, menunjukkan penambahan cluster lebih lanjut kurang signifikan. Dengan k=3, distribusi data adalah: Cluster 0 (487 data), Cluster 1 (35 data), dan Cluster 2 (125 data). Wordcloud digunakan untuk mengidentifikasi topik utama di setiap cluster.

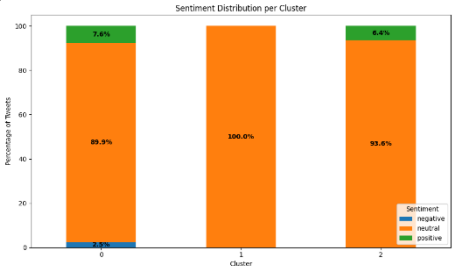
Gambar 10 Hasil Wordcloud

Berdasarkan ketiga *word cloud* yang Anda berikan, mari kita analisis masing-masing *cluster*. *Cluster* 0 kata-kata dominan adalah "karapan sapi", "Madura", "piala", "presiden", dan "budaya". Ini menunjukkan bahwa *cluster* ini sangat terkait dengan acara karapan sapi di Madura, termasuk aspek budaya, kompetisi, dan mungkin juga melibatkan tokoh-tokoh penting seperti presiden. *Cluster* 1 kata-kata dominan adalah kata-kata yang berkaitan dengan olahraga tradisional seperti "sepak takraw", "silat", "lompat", "panjat", dan "pencak silat". Ini mengindikasikan bahwa cluster ini berfokus pada berbagai jenis olahraga tradisional yang ada di Indonesia. *Cluster* 2 kata-kata dominan mirip dengan cluster 0, yaitu "karapan sapi", "piala", "Madura", dan juga nama-nama tokoh seperti "Farid Makruf" dan "Panglima". Namun, ada penambahan kata-kata seperti "pangdam", "Brawijaya", dan "TNI" yang mengindikasikan adanya keterlibatan pihak militer atau pemerintah dalam acara karapan sapi.

Tabel 1 Jumlah Sentimen setiap Cluster

Cluster	Sentimen Positif	Sentimen Netral	Sentimen Negatif
0	37	438	12
1	0	35	0
2	8	117	0

Seperti yang ditunjukkan oleh jumlah sentimen yang ditemukan untuk setiap *cluster* di Tabel 1, sentimen netral mendominasi semua *cluster*. Hal ini menunjukkan bahwa mayoritas opini tentang karapan sapi bersifat netral, mencerminkan penerimaan umum tanpa emosi yang kuat, baik positif maupun negatif.

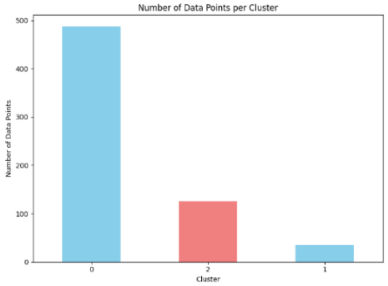


Gambar 11 Persentase Sentimen Setiap Cluster

Gambar 11 menunjukkan sentimen netral mendominasi semua cluster. Pada cluster 0, tweet netral (89.9%) terkait karapan sapi di Madura, dengan kata kunci seperti "karapan sapi", "Madura", dan "presiden", menunjukkan diskusi informatif terkait budaya dan kompetisi. Cluster 1 sepenuhnya netral (100%), fokus pada olahraga tradisional seperti "sepak takraw" dan "pencak silat". Pada cluster 2, sentimen netral (93.6%) terkait karapan sapi dan melibatkan tokoh militer atau pemerintah, dengan kata kunci seperti "Farid Makruf" dan "TNI". Semua cluster mencerminkan diskusi deskriptif tanpa emosi yang kuat.

4.5 Metode SMOTE

Teknik *resampling* sintetis minoritas (SMOTE) adalah solusi untuk masalah data tidak seimbang dengan menggunakan sampel sintetis dari kelas minoritas. Sehingga model pembelajaran mesin tidak bias terhadap kelas mayoritas, tujuan adalah untuk menyamakan distribusi kelas.



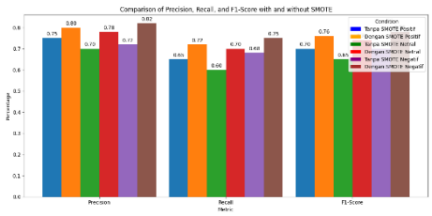
Gambar 12 Jumlah Sentimen Sebelum SMOTE

Gambar 12 menunjukkan distribusi data yang tidak seimbang. Cluster 0 mendominasi dengan hampir 500 data, diikuti cluster 2 dengan setengah jumlahnya, dan cluster 1 dengan data paling sedikit. Ketidakseimbangan ini dapat memengaruhi kinerja model, sehingga teknik seperti resampling mungkin diperlukan.

.

Gambar 13 Jumlah Sentimen Setelah SMOTE

Gambar 13 menunjukkan distribusi data sentimen setelah SMOTE, dengan kelas netral, positif, dan negatif masing-masing memiliki sekitar 450 data. SMOTE berhasil mengatasi ketidakseimbangan dengan membuat sampel sintetis untuk kelas minoritas, memastikan model lebih akurat dan tidak bias.

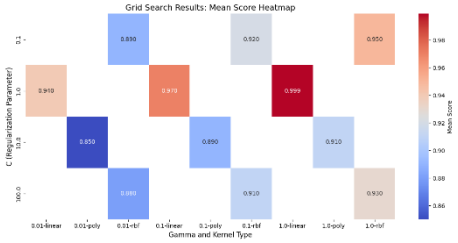


Gambar 14 Hasil Pengujian Pengaruh SMOTE

Gambar 14 menunjukkan bahwa penerapan SMOTE meningkatkan Precision, Recall, dan F1-Score pada semua kategori sentimen (positif, netral, negatif). Untuk sentimen negatif, peningkatan terbesar terlihat, dari Precision 0.78 menjadi 0.82, Recall 0.70 menjadi 0.75, dan F1-Score 0.75 menjadi 0.80. SMOTE secara konsisten memperbaiki kinerja model dengan menyeimbangkan data..

4.6 Klasifikasi SVM dan Evaluasi

Dalam tahap selanjutnya yakni pemilihan parameter terbaik pada SVM dengan menggunakan *Grid Search*.



Gambar 15 GridSearchCV

Gambar 15 merupakan *GridSearchCV* untuk mengevaluasi kombinasi parameter *C*, *gamma*, dan jenis kernel (linear, rbf, dan poly). Hasilnya divisualisasikan menggunakan heatmap yang menunjukkan performa model berdasarkan skor rata-rata dari validasi silang. Sumbu vertikal pada *heatmap* merepresentasikan nilai parameter regularisasi *C*, sedangkan sumbu horizontal menggambarkan kombinasi antara nilai gamma dan tipe kernel. Warna pada heatmap menunjukkan skor performa, di mana warna merah menunjukkan performa terbaik, dan biru menunjukkan skor yang lebih rendah.

Berdasarkan hasil *GridSearchCV*, parameter terbaik yang ditemukan adalah *C=1.0*, *gamma=1.0*, dan kernel linear, dengan skor rata-rata validasi sebesar 0.999, yang sangat mendekati nilai sempurna. Hal ini menunjukkan bahwa kernel linear dengan kombinasi parameter tersebut mampu menangkap pola dalam data secara optimal, menghasilkan model dengan performa terbaik. Dengan menggunakan parameter ini, model SVM dapat diandalkan untuk memprediksi data secara akurat. Visualisasi ini membantu memahami dampak kombinasi parameter terhadap performa model secara keseluruhan.

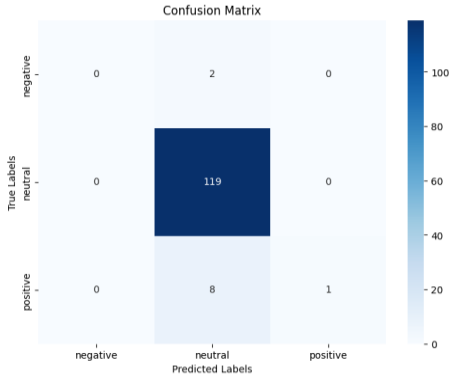
4.7 Evaluasi

Tabel 2 Laporan Klasifikasi

	Precision	Recall	F1 Score	Support
Negatif	0.00	0.00	0.00	2
Netral	0.92	1.00	0.96	119
Positif	1.00	0.11	0.20	9
Accuracy	0.92	0.92	0.92	130
Macro avg	0.64	0.37	0.39	130
Weighted avg	0.91	0.92	0.89	130

Berdasarkan Tabel 2, untuk kelas negative, *precision* mencapai 1.00 karena tidak ada prediksi salah positif, tetapi *recall-nya* 0.00 karena model gagal menangkap data yang benar-benar negatif. Untuk kelas neutral, *precision* dan *recall* sangat tinggi, masing-

masing 0.92 dan 1.00, menghasilkan *f1-score* sebesar 0.96. Namun, pada kelas positive, meskipun *precision* sempurna (1.00), *recall* sangat rendah (0.11), sehingga *f1-score* hanya 0.20. Hal ini menunjukkan model kesulitan dalam mengidentifikasi data positive. Secara keseluruhan, akurasi model mencapai 92%, namun nilai rata-rata *recall* pada semua kelas (macro avg) hanya 0.37, yang menunjukkan ketidakseimbangan performa model antar kelas. Kinerja tinggi pada kelas neutral mendominasi hasil, sementara kelas lainnya kurang terwakili. Hal ini mengindikasikan perlunya perbaikan, seperti penyeimbangan data atau pengoptimalan parameter model untuk menangani ketidakseimbangan kelas.



Gambar 16 Confusion Matrix

Dari *confusion matrix*, terlihat bahwa model sangat baik dalam memprediksi kelas neutral, dengan 119 prediksi yang benar dari total 119 data untuk kelas tersebut. Namun, untuk kelas negative dan positive, model menunjukkan kinerja yang kurang optimal. Tidak ada prediksi yang benar untuk negative, dan hanya 1 dari 9 data positive yang berhasil diklasifikasikan dengan benar. Sebagian besar data positive justru salah diprediksi sebagai neutral, mengindikasikan kesulitan model dalam membedakan kelas minoritas. Hasil ini menunjukkan bahwa meskipun model sangat baik dalam menangani kelas mayoritas (neutral), kinerja pada kelas minoritas (negative dan positive) masih harus ditingkatkan. Strategi seperti penyeimbangan data, pengaturan ulang parameter, atau penggunaan metode lain yang lebih sensitif terhadap ketidakseimbangan data dapat membantu meningkatkan performa model.

5. KESIMPULAN

Penelitian ini berhasil mengumpulkan 647 data ulasan aplikasi Twitter menggunakan teknik *crawling* berbasis *Python*. Dengan menggunakan metode K-Means, data ulasan dikelompokkan menjadi tiga *cluster* berdasarkan kesamaan fitur, dengan jumlah *cluster* optimal ditentukan melalui metode *Elbow*. Analisis *word cloud* menunjukkan bahwa *Cluster* 0 terkait dengan acara karapan sapi di Madura yang melibatkan unsur budaya, kompetisi, dan tokoh penting; *Cluster* 1 berfokus pada olahraga tradisional seperti pencak silat, sepak takraw, dan panjat; sedangkan *Cluster* 2 berkaitan dengan karapan sapi yang melibatkan pihak militer atau pemerintah.

Analisis sentimen menunjukkan bahwa sentimen netral mendominasi di ketiga *cluster*, mencerminkan bahwa sebagian besar diskusi bersifat informatif atau deskriptif, tanpa emosi yang signifikan. Ketidakseimbangan distribusi sentimen ini diatasi dengan penerapan teknik SMOTE, yang dapat mengimbangi jumlah data sentimen netral, negatif, dan positif. Teknik ini terbukti meningkatkan performa model pembelajaran mesin, khususnya pada kategori minoritas. Evaluasi model menggunakan SVM menunjukkan bahwa parameter terbaik ditemukan melalui *GridSearchCV*, yaitu kernel linear dengan nilai parameter *C=1.0* dan *gamma=1.0*, yang menghasilkan akurasi keseluruhan sebesar 92%. Namun, model masih menunjukkan ketidakseimbangan performa antar kelas, di mana kinerja pada kelas mayoritas (netral) sangat baik, tetapi model kesulitan menangani kelas minoritas (positif dan negatif), yang terlihat dari rendahnya nilai recall pada kedua kelas tersebut.

UCAPAN TERIMA KASIH

Penulis dengan penuh rasa terima kasih mengucapkan terima kasih kepada semua orang yang telah membantu menjalankan penelitian ini. Terutama, penulis mengucapkan terima kasih kepada pembimbing yang telah memberikan bimbingan, saran, dan dukungan yang luar biasa selama penelitian. Penulis juga mengucapkan terima kasih kepada rekan-rekan dan keluarga yang telah memberikan semangat dan inspirasi.

DAFTAR PUSTAKA

[1] M. Fachriza and H. Artikel, "Analisis Sentimen Kalimat Depresi Pada Pengguna Twitter Dengan Naive Bayes, Support Vector Machine, Random Forest," 2023. [Online]. Available: http://studentjournal.umpo.ac.id/index.php/komputek

[2] H. Fauzuna, "Makna Simbol Pada Upacara Kerapan Sapi Di Waru Pamekasan (Analisa Semiotika Roland Barthes)".

[3] W. Maulana Baihaqi, M. Pinilih, and M. Rohmah, "KOMBINASI K-MEANS DAN SUPPORT VECTOR MACHINE (SVM) UNTUK MEMPREDIKSI UNSUR SARA PADA TWEET," vol. 7, no. 3, 2020, doi: 10.25126/jtiik.202072126.

[4] R. Hermansyah and R. Sarno, "Sentiment analysis about product and service evaluation of pt telekomunikasi Indonesia tbk from tweets using textblob, naive bayes & K-NN Method," in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 511–516. doi: 10.1109/iSemantic50169.2020.9234238.

[5] S. Lestari, S. Saepudin, P. Studi, S. Informasi, F. Sains, and D. Teknologi, "Support Vector Machine: Analisis Sentimen Aplikasi Saham di Google Play Store," vol. 7, no. 2, pp. 81–90.

[6] F. S. Pamungkas and I. Kharisudin, "Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter," vol. 4, pp. 628–634, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[7] J. Teknik Komputer AMIK BSI, B. Rifai, S. Vambudi, R. Maulana, and F. Teknologi, "Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis Smote," vol. 8, no. 2, 2022, doi: 10.31294/jtk.v4i2.

[8] H. Harnelia, "ANALISIS SENTIMEN REVIEW SKINCARE SKINTIFIC DENGAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4095.

[9] Z. Ulfah Siregar, R. Ruli, A. Siregar, and R. Arianto, "KLASIFIKASI SENTIMENT ANALYSIS PADA KOMENTAR PESERTA DIKLAT MENGGUNAKAN METODE K-NEAREST NEIGHBOR," vol. 8, no. 1, 2019.

[10] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *Jurnal Sistem dan Teknologi Informasi (Justin)*, vol. 8, no. 2, p. 183, Apr. 2020, doi: 10.26418/justin.v8i2.36776.

[11] M. Nishom and D. S. Wibowo, "Implementasi Metode K-Means berbasis Chi-Square pada Sistem Pendukung Keputusan untuk Identifikasi Disparitas Kebutuhan Guru," *JURNAL SISTEM INFORMASI BISNIS*, vol. 8, no. 2, p. 187, Nov. 2018, doi: 10.21456/vol8iss2pp187-194.

[12] L. Luthfiana, J. Young Christian, and Rusli Andre, "Implementasi Algoritma Support Vector Machine dan Chi Square untuk Analisis Sentimen User Feedback Aplikasi," Nov. 2020.