

# ANALISIS SENTIMEN EVENT DI SIRKUIT MANDALIKA MENGUNAKAN ALGORITMA VADER-LEXICON DAN SVM BERBASIS PSO

Muhamad Rizal<sup>1</sup>

<sup>1, 2, 3</sup> Program Studi Ilmu Komputer, Universitas Bumigora  
Jl. Ismail Marzuki No.22, Cilinaya, Kec. Cakranegara, Kota Mataram, Nusa Tenggara Barat 83127  
e-mail: [muhamadrizal2nd@gmail.com](mailto:muhamadrizal2nd@gmail.com)<sup>1</sup>

(Naskah masuk : 8 Agustus 2025 Diterima untuk diterbitkan : 1 Mei 2026)

## ABSTRAK

Pariwisata merupakan sektor penting dalam peningkatan pendapatan dan citra Indonesia, terutama melalui penyelenggaraan event internasional. Penunjukan Indonesia sebagai tuan rumah MotoGP dan WSBK 2022 di Sirkuit Mandalika memberikan dampak besar bagi pemulihan ekonomi dan promosi pariwisata nasional. Namun, pelaksanaan kedua event tersebut masih menyisakan sejumlah kendala seperti fasilitas penunjang, infrastruktur, dan pengelolaan lapangan yang menimbulkan banyak opini publik di media sosial. Fenomena ini menunjukkan pentingnya pemetaan sentimen masyarakat untuk mengevaluasi keberhasilan event dan memahami persepsi publik secara objektif. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terkait event di Sirkuit Mandalika menggunakan algoritma Support Vector Machine (SVM) berbasis Particle Swarm Optimization (PSO), dengan pelabelan otomatis VADER-Lexicon. Penelitian mengikuti tahapan CRISP-DM mulai dari pengumpulan data hingga evaluasi. Sebanyak 7.184 data dikumpulkan melalui web scraping dari platform X dan YouTube. Setelah melalui preprocessing dan ekstraksi fitur TF-IDF, dilakukan klasifikasi menggunakan SVM dan SVM-PSO dengan dua skenario: dua sentimen (positif-negatif) dan tiga sentimen (positif-negatif-netral). Hasil pengujian menunjukkan bahwa SVM-PSO memberikan akurasi tertinggi pada skenario dua sentimen sebesar 92,17% dan memiliki peningkatan akurasi 11,34% dibandingkan tiga sentimen. Temuan ini menunjukkan bahwa optimasi PSO mampu meningkatkan performa SVM dalam analisis sentimen. Namun, variasi bahasa informal pada media sosial menjadi salah satu keterbatasan penelitian karena dapat mempengaruhi kualitas preprocessing dan akurasi model.

**Kata Kunci:** Pariwisata, Media Sosial, Event MotoGP dan WSBK di Sirkuit Mandalika, Analisis Sentimen, VADER-Lexicon, Support Vector Machine, Particle Swarm Optimization.

## ABSTRACT

Tourism is a vital sector in enhancing Indonesia's revenue and image, particularly through the organization of international events. Indonesia's appointment as the host of MotoGP and WSBK 2022 at the Mandalika Circuit has had a significant impact on economic recovery and national tourism promotion. However, the implementation of these events still faced several challenges, such as supporting facilities, infrastructure, and field management, which triggered numerous public opinions on social media. This phenomenon highlights the importance of mapping public sentiment to evaluate event success and understand public perceptions objectively. This study aims to analyze public sentiment regarding the events at the Mandalika Circuit using the Support Vector Machine (SVM) algorithm optimized with Particle Swarm Optimization (PSO), employing automatic labeling through the VADER-Lexicon. The research follows the CRISP-DM stages, from data collection to evaluation. A total of 7,184 data points were collected through web scraping from the X and YouTube platforms. After preprocessing and feature extraction using TF-IDF, classification was performed using SVM and SVM-PSO under two scenarios: two sentiments (positive-negative) and three sentiments (positive-negative-neutral). The experimental results show that SVM-PSO achieved the highest accuracy in the two-sentiment scenario at 92.17%, with an accuracy improvement of 11.34% compared to the three-sentiment scenario. These findings indicate that PSO optimization effectively enhances SVM performance in sentiment analysis. However, variations in informal language on social media remain a limitation of this study, as they may affect preprocessing quality and model accuracy.

**Keywords:** Tourism, Social Media, MotoGP and WSBK Events at Mandalika Circuit, Sentiment Analysis, VADER-Lexicon, Support Vector Machine, Particle Swarm Optimization.

## I. PENDAHULUAN

Pariwisata merupakan salah satu sektor strategis yang berperan penting dalam peningkatan pendapatan negara serta pemulihan ekonomi setelah pandemi COVID-19. Indonesia memiliki potensi besar karena keanekaragaman budaya, kekayaan alam, dan destinasi berskala internasional [1]. Salah satu pengembangan pariwisata nasional yang menarik perhatian dunia adalah penyelenggaraan event balap internasional MotoGP dan WSBK di Sirkuit Mandalika, Nusa Tenggara Barat. Agenda ini tidak hanya memperkuat brand pariwisata Indonesia, tetapi juga diharapkan menjadi pendorong investasi serta pertumbuhan ekonomi daerah [2][3].

Walaupun event MotoGP dan WSBK berjalan dengan sukses, berbagai permasalahan masih muncul, terutama terkait sarana dan prasarana seperti pengelolaan transportasi, fasilitas pendukung, infrastruktur jalan, hingga pelayanan wisatawan [4]. Permasalahan tersebut memicu meningkatnya opini publik di media sosial. Media sosial seperti X (Twitter) dan YouTube menjadi ruang diskusi yang sangat aktif, di mana masyarakat menyampaikan dukungan, kritik, maupun keluhan. Opini publik ini dapat bersifat positif, negatif, atau netral, sehingga dibutuhkan metode analisis sentimen untuk mengetahui bagaimana persepsi masyarakat terhadap penyelenggaraan event internasional tersebut.

Analisis sentimen merupakan salah satu cabang *text mining* yang digunakan untuk mengidentifikasi emosi atau opini publik dalam bentuk teks, baik positif, negatif, maupun netral [5]. Data dari media sosial menjadi sumber informasi yang penting bagi pemerintah dan pengelola kawasan wisata, karena dapat digunakan untuk mengukur kepuasan, mengevaluasi kualitas layanan, serta menentukan strategi peningkatan fasilitas [6]. Seiring meningkatnya jumlah pengguna internet dan media sosial, metode manual sudah tidak efektif, sehingga diperlukan algoritma kecerdasan buatan untuk menganalisis data dalam skala besar.

Berbagai penelitian terkait analisis sentimen pariwisata telah dilakukan dalam beberapa tahun terakhir. Penelitian [5] menggunakan metode *Multinomial Naïve Bayes Classifier* dengan 6.148 data dan memperoleh akurasi sebesar 78%. Selanjutnya, penelitian [6] menggunakan metode *Support Vector Machine (SVM)* pada 1.000 data dan berhasil mencapai akurasi 90%. Adapun penelitian [7] membandingkan beberapa metode, yaitu *Naïve Bayes*, *Support Vector Machine*, serta versi masing-masing yang dioptimasi menggunakan *Particle Swarm Optimization (PSO)*, pada analisis sentimen artikel berita tokoh sepak bola dunia. Hasilnya menunjukkan bahwa metode *SVM berbasis PSO* memberikan kinerja terbaik dengan akurasi 84,00%, diikuti oleh *NB PSO* sebesar 83,00%, *SVM* sebesar 78,50%, dan *NB* sebesar 76,50%. Namun, jumlah data yang digunakan pada penelitian tersebut masih terbatas, yaitu hanya 100 data.

Namun, penelitian sebelumnya pada kasus Mandalika masih terbatas pada jumlah data, metode tunggal, dan hanya menggunakan satu sumber *platform* komentar. Penelitian [7] menggunakan 6.184 data dan hanya memakai algoritma *Multinomial Naïve Bayes*. Selain itu, belum banyak studi yang membandingkan performa akurasi antara klasifikasi dua sentimen (positif-negatif) dan tiga sentimen (positif-negatif-netral) pada kasus *event* MotoGP dan WSBK. Padahal, kategori netral sering mendominasi komentar publik dan berpotensi mempengaruhi hasil model.

Berdasarkan *research gap* tersebut, penelitian ini mengusulkan analisis sentimen komentar masyarakat mengenai event MotoGP dan WSBK di Sirkuit Mandalika menggunakan metode *Support Vector Machine berbasis Particle Swarm Optimization (SVM-PSO)* dan menggunakan *VADER-Lexicon* untuk proses *labelling* otomatis karena mampu mengolah teks pendek, slang, hingga *emoticon* dengan baik [8]. Jumlah data yang digunakan sebanyak 7.184 komentar yang diambil dari *platform* X dan YouTube, lebih banyak dibandingkan penelitian sebelumnya sehingga diharapkan menghasilkan model yang lebih stabil. Selain itu, penelitian ini membandingkan performa akurasi pada dua sentimen dan tiga sentimen untuk mengetahui konfigurasi terbaik. Penelitian ini diharapkan memberikan kontribusi pada peningkatan kualitas layanan dan fasilitas Sirkuit Mandalika berdasarkan opini masyarakat berbasis data. Hasil analisis dapat digunakan oleh pemerintah daerah, pengelola kawasan ekonomi khusus (KEK Mandalika), dan pemangku kebijakan sebagai rekomendasi pengembangan pariwisata berkelanjutan di Nusa Tenggara Barat.

## II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Cross Industry Standard Process for Data Mining (CRISP-DM)*, yang merupakan standar industri dalam pengembangan proyek data mining. Metodologi ini terdiri dari lima tahapan utama: *Business Understanding*, *Data Understanding*, *Data Preparation*,

*Modeling*, dan *Evaluation*. Setiap tahap saling terkait dan dapat dilakukan secara iteratif sesuai kebutuhan.

### 2.1 Business Understanding

Pada tahap *business understanding* bertujuan untuk memahami bidang masalah, menghasilkan solusi yang tepat, dan mengungkapkan faktor penting yang berpengaruh pada hasil penelitian. Pada penelitian ini akan melakukan analisis sentimen komentar mengenai *event* besar yang telah berlangsung di Sirkuit Mandalika yaitu *MotoGP* dan *WSBK 2022* pada media sosial *x* dan *youtube*. Komentar pada media sosial *x* dan *youtube* sangat beragam dalam mengekspresikan berbagai macam pendapat baik positif, negatif atau netral pada kolom komentar yang disediakan.

### 2.2 Data Understanding

Tahap ini bertujuan untuk mengumpulkan, mengidentifikasi, dan memahami data yang digunakan dalam penelitian. Total data yang dikumpulkan sebanyak 7.184 data mentah dari dua sumber media sosial, yaitu Media Sosial *X* dan *YouTube*, yang berisi komentar masyarakat terkait *event MotoGP* dan *WSBK 2022* di Sirkuit Mandalika. Data yang dikumpulkan mencakup tiga jenis sentimen, yaitu positif, negatif, dan netral.

Pengumpulan data dari Media Sosial *X* dilakukan dengan metode *crawling* menggunakan *RapidMiner*, yang mencakup tahapan seperti mengambil data menggunakan API dengan kata kunci “sirkuitmandalika”, memilih atribut teks, menghapus duplikat, dan menyimpannya ke dalam *format Excel*. Proses ini dilakukan selama dua hari, yaitu pada tanggal 10–11 Januari 2023.

Sementara itu, pengumpulan data dari *YouTube* dilakukan dengan metode *web scraping* menggunakan *JavaScript* dan *Node.js* dengan bantuan *library Puppeteer*. Proses *scraping* mencakup pengambilan *link* video, navigasi ke halaman, identifikasi elemen komentar, *scroll* otomatis, dan ekstraksi komentar yang kemudian disimpan dalam *format CSV*. *Scraping* ini dilakukan pada tanggal 14 Maret 2023.

### 2.3 Data Preparation

Tahap ini bertujuan untuk mempersiapkan data mentah hasil *web scraping* dari Media Sosial *X* dan *YouTube* agar siap dianalisis. Proses ini dilakukan melalui serangkaian tahap *preprocessing* yang penting untuk membersihkan dan menyusun data secara sistematis. *Text Preprocessing* adalah tahap awal pengolahan teks sebelum dianalisis, meliputi *cleaning*, *case folding*, *normalization*, *filtering*, *tokenizing*, *stopword removal*, *stemming*, dan *labeling* [9]. Langkah pertama adalah *cleaning* dan *case folding*, yaitu menghapus karakter tidak perlu dan mengubah semua huruf menjadi huruf kecil. Selanjutnya dilakukan *normalization* untuk mengganti kata tidak baku (*slang*) menjadi kata yang sesuai berdasarkan kamus. *Normalization* dilakukan menggunakan kamus slang Bahasa Indonesia buatan peneliti yang disimpan dalam file *dictslang1.txt*. Kamus ini berisi pasangan kata slang dan kata baku dalam format *slang|baku*, misalnya: “*gk|tidak*”, “*bgt|banget*”, “*gmna|bagaimana*”. Kamus ini di-load secara otomatis dan digunakan mengganti seluruh kata slang menjadi bentuk baku. Kamus ini disusun berdasarkan kumpulan kata tidak baku yang umum digunakan di media sosial Indonesia, kemudian diverifikasi dengan padanan kata baku dari KBBI daring. *Filtering* digunakan untuk membuang kata-kata yang kurang penting menggunakan metode *stoplist*, diikuti oleh *tokenizing* untuk memecah kalimat menjadi kata-kata terpisah.

Setelah itu dilakukan *stopword removal*, yakni penghapusan kata-kata umum yang tidak memiliki makna penting dalam analisis seperti “yang”, “dan”, “di”. Tahap berikutnya adalah *stemming*, yang mengubah setiap kata menjadi bentuk dasarnya (*root word*). Terakhir, dilakukan *labeling* menggunakan *VADER-Lexicon* untuk mengklasifikasikan sentimen dari setiap komentar menjadi positif, negatif, atau netral. *Labeling* dilakukan dalam dua skema, yaitu tiga kelas sentimen (positif, netral, dan negatif) serta dua kelas sentimen (positif dan negatif). Pada skema tiga kelas, dari total 6.530 data hasil *preprocessing*, diperoleh 3.176 data positif, 2.365 data netral, dan 989 data negatif. Sedangkan pada skema dua kelas diperoleh 5.509 data positif dan 1.021 data negatif. Komposisi data ini menunjukkan dominasi sentimen positif pada komentar yang dianalisis, serta menjadi dasar dalam tahap pelatihan model klasifikasi pada tahap modeling berikutnya.

*VADER-Lexicon* adalah metode *lexicon-based sentiment analysis* yang menggunakan kamus kata dengan skor valensi untuk menentukan polaritas positif, negatif, atau netral berdasarkan *compound score* [10][8]. Berikut rumus mencari *compound VADER-Lexicon*:

$$\text{compound} = \frac{\sum \text{valensi kata}}{\sqrt{\sum (\text{valensi kata})^2 + \alpha}}$$

Dengan:

- $\sum \text{valensi kata} = \text{total skor valensi semua kata dalam kalimat}$
- $\sum (\text{valensi kata})^2 = \text{jumlah kuadrat valensi setiap kata}$
- $\alpha = \text{nilai normalisasi bernilai 15 (konstanta dalam formula VADER)}$

Hasil *compound* dibagi menjadi tiga kelas:

- $\geq 0.05 \rightarrow \text{positif}$
- $\leq -0.05 \rightarrow \text{negatif}$
- di antara keduanya  $\rightarrow \text{netral}$

Pendekatan ini cocok untuk komentar pendek dan informal seperti komentar media sosial. Keunggulannya adalah tidak memerlukan data berlabel, sehingga cocok digunakan pada data komentar media sosial yang besar. Seluruh alur *preprocessing* tersebut dapat dilihat pada Gambar 1, yang memvisualisasikan urutan dan keterkaitan antar proses dalam tahap pembersihan data.



Gambar 1. Tahap *Preprocessing*

## 2.4 Modeling

Tahap *modeling* dilakukan untuk membangun model klasifikasi pada dataset komentar yang telah melewati proses *preprocessing*. Algoritma yang digunakan adalah *Support Vector Machine* (SVM), dengan peningkatan performa menggunakan *Particle Swarm Optimization* (PSO). Sebelum klasifikasi, dilakukan ekstraksi fitur menggunakan metode TF-IDF. TF-IDF (*Term Frequency-Inverse Document Frequency*) digunakan sebagai metode pembobotan kata dalam dokumen. Prosesnya dimulai dengan menghitung frekuensi setiap kata (*Term Frequency*), lalu menghitung pembobotannya dengan mempertimbangkan seberapa jarang kata tersebut muncul di seluruh dokumen (*Inverse Document Frequency*) [11].

**Rumus:**

$$TF - IDF = TF \times IDF \quad (1)$$

$$TF = \frac{\text{jumlah frekuensi kata terpisah}}{\text{jumlah kata}} \quad (2)$$

$$IDF = \log \left( \frac{\text{jumlah dokumen}}{\text{jumlah frekuensi kata terpisah}} \right) \quad (3)$$

$$IDF = \log \left( \frac{7}{DF} \right) \quad (4)$$

**Keterangan:**

- *TF - IDF* memberikan bobot pentingnya suatu kata dalam dokumen.
- *TF* mengukur seberapa sering kata muncul dalam dokumen.
- *IDF* mengukur kelangkaan kata di seluruh dokumen (korpus).

Setelah ekstraksi fitur, dilakukan klasifikasi menggunakan SVM berbasis PSO. Tahapan awal yaitu klasifikasi menggunakan SVM, kemudian evaluasi model menggunakan *K-Fold Cross Validation*, dan terakhir optimasi menggunakan PSO. Pada penelitian ini, proses pembagian data dilakukan sebelum tahap klasifikasi. Data dibagi menggunakan metode *train-test split* dengan proporsi 80% sebagai data latih (*training set*) dan 20% sebagai data uji (*testing set*). Pembagian ini dilakukan untuk memperoleh model awal dan menghindari evaluasi yang hanya bergantung pada satu set data.

Selain itu, penelitian ini juga menerapkan metode *10-Fold Cross Validation* untuk proses evaluasi lanjutan. Pada setiap *fold*, 90% data digunakan sebagai data latih dan 10% sebagai data uji. Proses TF-IDF diterapkan hanya pada data latih pada setiap *fold* untuk mencegah data *leakage*, sedangkan data uji hanya ditransformasikan menggunakan matriks TF-IDF dari data latih. Penggunaan kedua metode ini (*train-test split* dan *K-Fold*) memungkinkan model dievaluasi secara lebih komprehensif dan memastikan kemampuan generalisasi yang lebih baik.

*Support Vector Machine* (SVM) bekerja dengan mencari *hyperplane* terbaik sebagai pemisah antara dua kelas data. SVM memetakan data ke ruang berdimensi tinggi agar bisa dipisahkan meskipun tidak *linear* [12].

**Rumus SVM:**

$$z = \omega t \times x + b \quad (5)$$

**Keterangan:**

- $z$  : skor prediksi.
- $\omega$  : vektor bobot.
- $x$  : vektor fitur (hasil dari TF-IDF).
- $b$  : bias (nilai *offset* terhadap *hyperplane*).

*K-Fold Cross Validation* digunakan untuk mengevaluasi model dengan membagi data ke dalam 10 bagian (fold). Setiap bagian secara bergiliran menjadi data uji dan sisanya menjadi data latih. Evaluasi dilakukan selama 10 kali iterasi, lalu hasilnya dirata-rata untuk mendapatkan performa model yang stabil dan akurat. *Particle Swarm Optimization* (PSO) adalah algoritma optimasi berbasis populasi, di mana setiap partikel mewakili solusi dan bergerak dalam ruang pencarian untuk menemukan parameter terbaik bagi SVM [13]. PSO dapat digunakan untuk optimasi parameter dan seleksi fitur [14]. Langkah-langkah PSO:

1. Inisialisasi partikel (jumlah, posisi, kecepatan).
2. Evaluasi performa SVM pada tiap partikel.
3. Update posisi dan kecepatan menggunakan rumus.

**Rumus:**

$$vi(t + 1) = wvi(t) + c1r1(pibest - xi(t)) + c2r2(gbest - xi(t)) \quad (6)$$

$$xi(t + 1) = xi(t) + vi(t + 1) \quad (7)$$

**Keterangan:**

- $vi$  : kecepatan partikel ke-i.
- $xi$  : posisi partikel ke-i.
- $pibest$  : posisi terbaik lokal partikel.
- $gbest$  : posisi terbaik global.
- $w$  : faktor inersia.
- $c1, c2$  : koefisien akselerasi.
- $r1, r2$  : bilangan acak antara 0 dan 1.

Dengan memanfaatkan PSO, model klasifikasi SVM dapat dioptimalkan lebih baik, baik dari segi akurasi maupun efisiensi.

## 2.5 Evaluasi

Tahap ini bertujuan mengevaluasi hasil analisis sentimen yang dilakukan menggunakan *confusion matrix*. *Confusion matrix* menunjukkan seberapa baik model klasifikasi dalam memprediksi kelas target, dan divisualisasikan dalam bentuk *heatmap* agar lebih mudah dipahami. Model dievaluasi berdasarkan *accuracy* serta metrik lain seperti *precision*, *recall*, dan *F1-score*. Evaluasi dilakukan terhadap performa model setelah proses pelatihan (*training*) dan pengujian (*testing*) dengan mempertimbangkan rata-rata makro, mikro, dan skor F1 tertimbang. Hal ini berguna untuk menangani data dengan dua kelas atau lebih. Jika performa model belum optimal, maka dilakukan perbaikan pada tahapan sebelumnya.

**Rumus evaluasi untuk dua kelas:**

$$accuracy = \frac{TP}{\text{jumlah data}} \quad (8)$$

$$precision = \frac{(TP+TP)/(TP+FP)A+(TP+FP)B}{2} \quad (9)$$

$$recall = \frac{(TP+TP)/(TP+FN)A+(TP+FN)B}{2} \quad (10)$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (11)$$

**Rumus evaluasi untuk tiga kelas:**

$$accuracy = \frac{TP}{\text{jumlah data}} \quad (12)$$

$$precision = \frac{(TP+TP)/(TP+FP)A+(TP+FP)B+(TP+FP)C}{3} \quad (13)$$

$$recall = \frac{(TP+TP)/(TP+FN)A+(TP+FN)B+(TP+FN)C}{3} \tag{14}$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{15}$$

**Keterangan:**

- *TP (True Positive)*: Prediksi benar terhadap data yang benar.
- *TN (True Negative)*: Prediksi salah terhadap data yang memang salah.
- *FP (False Positive)*: Prediksi benar padahal kenyataannya salah.
- *FN (False Negative)*: Prediksi salah padahal kenyataannya benar [15].

**III. HASIL DAN PEMBAHASAN**

Bab ini membahas tahapan dan hasil penerapan analisis sentimen terhadap komentar masyarakat terkait *event MotoGP* dan *WSBK 2022* di Sirkuit Mandalika. Analisis dilakukan mengikuti tahapan CRISP-DM: *business understanding, data understanding, data preparation, modeling, dan evaluation*. Pada tahap *business understanding*, dilakukan pemahaman terhadap objek penelitian dengan mengumpulkan komentar dari media sosial X dan YouTube yang berisi pendapat positif maupun negatif. Selain itu, tahap ini juga mencakup penentuan metode terbaik untuk analisis sentimen, yaitu *Support Vector Machine* yang dioptimalkan dengan *Particle Swarm Optimization*.

Selanjutnya, dilakukan tahapan data *understanding* untuk mengenal karakteristik data yang akan digunakan. Pengumpulan data dilakukan dari dua sumber utama, yaitu media sosial X (*Twitter*) dan *YouTube*. Pada media sosial X, data diperoleh melalui proses *crawling* sebanyak dua kali pada tanggal 10 dan 11 Januari 2023, dengan hasil sebanyak 1.193 dan 1.329 data komentar. Sementara pada *platform YouTube*, data diperoleh melalui metode *web scraping* yang dilakukan pada tanggal 14 Maret 2023 dan menghasilkan sebanyak 4.662 data komentar. Data yang diperoleh dari kedua *platform* memiliki bentuk komentar teks berbahasa Indonesia yang mengandung opini terhadap suatu isu.

Tahap data *preparation* merupakan proses penyiapan data agar siap digunakan dalam pemodelan dan analisis. Data yang diperoleh dari hasil *crawling* media sosial X dan *scraping YouTube* terlebih dahulu digabungkan menjadi satu *dataset*. Penggabungan ini menghasilkan total 7.184 data komentar dari kedua *platform*, sebagaimana ditunjukkan pada Tabel 1.

**Tabel 1. Hasil Gabungan Data Media Sosial X dan Youtube**

No	Text
1	Bahagia banget dek....
2	Polda NTB Antisipasi Keramaian Lalu Lintas Di MotoGP Mandalika
3	Pagelaran #WSBK2022 Indonesia akan dimeriahkan penampilan artis nasional
4	Ditonton Ratusan Juta Orang Di Dunia
5	Siapa yang akan jadi pemenang?

Setelah data digabungkan, dilakukan tahapan *preprocessing* secara bertahap untuk membersihkan dan menyiapkan data mentah menjadi data yang berkualitas. Langkah pertama adalah proses *cleaning* dan *case folding*. *Cleaning* dilakukan dengan menghapus unsur-unsur yang tidak relevan dalam analisis seperti *mention (@), hashtag (#), link URL*, simbol-simbol khusus, angka, serta kata-kata yang tidak memiliki makna penting dalam konteks sentimen. Kemudian dilakukan *case folding*, yaitu proses mengubah seluruh huruf menjadi huruf kecil agar format teks menjadi seragam. Contoh hasil *cleaning* dan *case folding* dapat dilihat pada Tabel 2.

**Tabel 2. Hasil Cleaning dan Case Folding Data**

	Text	Clean_text
0	Bahagia banget dek....	bahagia banget dek
1	Polda NTB Antisipasi Keramaian Lalu Lintas Di MotoGP Mandalika	polda ntb antisipasi keramaian lalu lintas di motogp mandalika

2	Pagelaran #WSBK2022 Indonesia akan dimeriahkan penampilan artis nasional	pagelaran indonesia akan dimeriahkan penampilan artis nasional
3	Ditonton Ratusan Juta Orang Di Dunia	ditonton ratusan juta orang di dunia
4	Siapa yang akan jadi pemenang?	Siapa yang akan jadi pemenang

Selanjutnya dilakukan proses normalisasi, yaitu memperbaiki kata-kata yang salah eja atau disingkat menjadi bentuk baku, misalnya kata “bgt” menjadi “banget” atau “adik” menggantikan “dek”. Normalisasi ini membantu menyatukan variasi kata ke dalam bentuk standar. Tahap berikutnya adalah *filtering*, yaitu proses menghapus data ganda atau duplikat (data *dummy*) agar tidak terjadi bias dalam pemodelan. Dari 7.184 data awal, sebanyak 605 data teridentifikasi sebagai data ganda dan dihapus, sehingga diperoleh 6.580 data bersih yang siap untuk diproses lebih lanjut (Tabel 3).

**Tabel 3. Hasil Filter Data Dummy / Data Kembar**

Data Terfilter: 6.580

Data Terhapus: 605

		<i>Text</i>
0		wsbk mandalika jadi penentu juara lagi
1	cobain asli tetapi palsu sirkuit mandalika pakai honda civic type r	
2	terus keberhasil balapan sirkuit mandalika karena keberhasilan anies begitu	
3	harusnya malu orang yang dulu tidak suka dengan formula e kini setelah anis tidak jadi gubernur dki mereka sadar bahwa formula e lebih bermanfaat daripada sirkuit mandalika	
4		kembali ke kepercayaan masing sih

Data yang telah difilter kemudian diproses dengan *tokenizing*, yaitu memecah kalimat atau teks menjadi bagian-bagian kecil (*token*) berupa kata-kata terpisah. *Token-token* yang diperoleh kemudian melalui proses *stopword removal*. Proses ini bertujuan membuang kata-kata umum dalam bahasa Indonesia yang tidak memiliki nilai informasi signifikan, seperti “yang”, “di”, “dan”, atau “dari”. Dengan menghilangkan kata-kata tersebut, hanya *term-term* penting yang mewakili inti opini yang dipertahankan. Setelah *stopword removal*, dilakukan tahap *stemming* untuk mengubah kata berimbuhan menjadi bentuk dasar. Misalnya, “menonton” menjadi “tonton”, “kemacetan” menjadi “macet”, dan sebagainya. *Stemming* membantu dalam menyamakan berbagai variasi kata ke akar katanya agar lebih konsisten untuk analisis sentimen. Hasil *stemming* dapat dilihat pada Tabel 4.

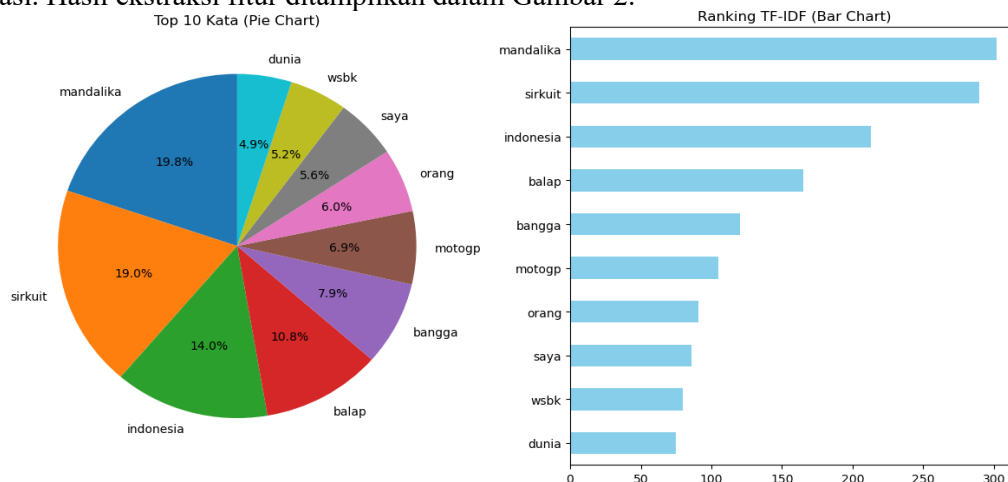
**Tabel 4. Hasil Stemming Data**

	<i>Removed_text</i>	<i>Stemmed_text</i>	<i>Stemmed_textjoin</i>
0	[sangat]	[bahagia, adik]	bahagia adik
1	[lalu, di]	[polda, ntb, antisipasi, macet, lintas, motogp, mandalika]	polda ntb antisipasi macet lintas motogp mandalika
2	[akan, oleh]	[tunjuk, indonesia, riah, tampil, artis, nasional]	tunjuk indonesia riah tampil artis nasional
3	[oleh, di, seluruh]	[tonton, ratus, juta, orang, dunia]	tonton ratus juta orang dunia
4	[siapa, yang, akan, menjadi]	[menang]	menang

Tahap terakhir dalam data *preparation* adalah *labeling*, yaitu proses pemberian label sentimen terhadap masing-masing data berdasarkan skor yang dihasilkan menggunakan metode *VADER-Lexicon*. *Labeling* dilakukan untuk dua skema: tiga kelas sentimen (positif, netral, dan negatif) dan dua kelas sentimen (positif dan negatif). Hasilnya, untuk tiga kelas diperoleh 3.176 data positif, 2.365 netral, dan 989 negatif. Sementara itu, untuk dua kelas diperoleh 5.509 data positif dan 1.021 data negatif.

Setelah data siap digunakan, dilakukan tahap *modeling* untuk membangun model klasifikasi sentimen berdasarkan data yang telah diproses. *Modeling* merupakan tahap pemilihan teknik data *mining* dengan menentukan algoritma yang digunakan untuk proses klasifikasi data. Pada penelitian ini, digunakan metode *Support Vector Machine* (SVM) yang dioptimalkan menggunakan *Particle Swarm Optimization* (PSO). Tahapannya dimulai dari proses ekstraksi fitur menggunakan TF-IDF, dilanjutkan dengan klasifikasi menggunakan SVM, kemudian dilakukan evaluasi model menggunakan *10-Fold Cross Validation*, dan terakhir dilakukan optimasi parameter menggunakan PSO. Setelah data siap digunakan, dilakukan pembagian dataset menggunakan *train-test split* dengan proporsi 80% sebagai data latih dan 20% sebagai data uji untuk membangun model awal. Selanjutnya, proses validasi model dilakukan menggunakan metode *10-Fold Cross Validation* guna memastikan kestabilan dan kemampuan generalisasi model.

Ekstraksi fitur dilakukan menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengubah data teks menjadi bentuk numerik yang dapat diproses oleh algoritma klasifikasi. Hasil ekstraksi fitur ditampilkan dalam Gambar 2.



Gambar 2. Hasil Ekstraksi Fitur TD-IDF

Setelah itu, data hasil ekstraksi diklasifikasikan menggunakan metode *Support Vector Machine* (SVM). Hasil klasifikasi menunjukkan bahwa akurasi klasifikasi dua sentimen lebih tinggi dibandingkan dengan tiga sentimen. Adapun hasil klasifikasi dengan metode SVM untuk dua sentimen 0,89 dan tiga sentimen 0,80. Untuk mengevaluasi performa model secara menyeluruh, digunakan teknik *10-Fold Cross Validation*. Validasi ini bertujuan untuk memastikan bahwa model yang dikembangkan bersifat general dan tidak hanya cocok untuk data latih. Hasil evaluasi pada data dua dan tiga sentimen ditampilkan pada Tabel 5. dengan rincian sebagai berikut:

Tabel 5. Hasil *K-Fold Cross Validation*

c	Dua Sentimen	Tiga Sentimen
Akurasi	0,89	0,79
Presisi	0,88	0,80
Recall	0,89	0,79
F1-Score	0,88	0,79

Tahap selanjutnya adalah optimasi parameter SVM menggunakan *Particle Swarm Optimization* (PSO). Pada tahap ini, proses pencarian nilai *hyperparameter C* terbaik dilakukan untuk meningkatkan kinerja model. PSO dijalankan hingga maksimal 50 iterasi. Hasil klasifikasi setelah optimasi menggunakan SVM berbasis PSO ditampilkan pada Tabel 6. Hasilnya menunjukkan peningkatan akurasi, khususnya pada data dua sentimen:

**Tabel 6. Klasifikasi Metode SVM PSO Dua dan Tiga Sentimen**

*Stopping search: maximum iterations reached - -> 50*

*Hyperparameter terbaik (c) untuk dua sentimen yang dioptimalkan: 3,66*

*Hyperparameter terbaik (c) untuk tiga sentimen yang dioptimalkan: 1,43*

Sentimen	Hyperparameter C	Akurasi
Dua Sentimen	3,66	0,92
Tiga Sentimen	1,43	0,80

Dengan demikian, dapat disimpulkan bahwa penggunaan PSO dalam proses optimasi parameter pada SVM mampu meningkatkan akurasi klasifikasi, terutama untuk klasifikasi dua sentimen. Terakhir tahap *evaluation*. Hasil evaluasi dilakukan dengan menggunakan *confusion matrix* untuk mengukur performa model klasifikasi, baik pada metode SVM maupun SVM yang dioptimasi dengan *Particle Swarm Optimization* (PSO). Evaluasi ini menghasilkan metrik performa berupa akurasi, presisi, *recall*, dan *f1-score*. Evaluasi dilakukan pada dua jenis data, yaitu dua sentimen dan tiga sentimen. Berdasarkan hasil evaluasi, metode SVM berbasis PSO menunjukkan performa terbaik, khususnya pada klasifikasi dua sentimen, yang memperoleh nilai tertinggi pada seluruh metrik evaluasi dibandingkan metode lainnya. Sementara itu, pada klasifikasi tiga sentimen, metode SVM berbasis PSO juga sedikit lebih unggul dibandingkan SVM biasa, meskipun selisihnya tidak terlalu besar. Rekapitulasi lengkap hasil evaluasi pada masing-masing metode dan jumlah sentimen dapat dilihat pada Tabel 7. berikut:

**Tabel 7. Hasil Evaluasi SVM dan SVM + PSO Dua dan Tiga Sentimen**

Metode	Jumlah Sentimen	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)
SVM	Dua Sentimen	89,88	85,89	70,81	77,62
SVM	Tiga Sentimen	80,15	78,67	75,00	76,79
<b>SVM + PSO</b>	<b>Dua Sentimen</b>	<b>92,17</b>	<b>83,54</b>	<b>81,20</b>	<b>82,36</b>
SVM + PSO	Tiga Sentimen	80,82	79,28	75,63	77,41

#### IV. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian analisis sentimen pada *event* besar di Sirkuit Mandalika menggunakan algoritma *Vader-Lexicon* dan SVM berbasis PSO, dengan data yang diklasifikasikan ke dalam dua sentimen (positif dan negatif) serta tiga sentimen (positif, negatif, dan netral), ditemukan adanya perbedaan akurasi sebesar 11,34% pada klasifikasi menggunakan SVM PSO antara dua dan tiga sentimen, serta 9,73% pada klasifikasi menggunakan SVM tanpa PSO. Akurasi tertinggi, yakni 92,17%, diperoleh saat menggunakan SVM PSO pada data dengan dua sentimen. Hasil ini menunjukkan bahwa penggunaan PSO mampu meningkatkan kinerja SVM dalam analisis sentimen, khususnya dalam konteks *event* besar di Sirkuit Mandalika. Dengan demikian, dapat disimpulkan bahwa SVM PSO lebih unggul dibandingkan SVM konvensional, dan bahwa klasifikasi dengan dua sentimen memberikan akurasi yang lebih baik dibandingkan tiga sentimen.

Saran untuk penelitian selanjutnya yaitu melakukan pengolahan data yang lebih baik guna meningkatkan akurasi hasil analisis sentimen. Selain itu, disarankan untuk mencoba metode lain di luar SVM dan PSO untuk perbandingan hasil analisis, memperluas lingkup penelitian dengan melibatkan lebih banyak responden dari berbagai latar belakang, serta mempertimbangkan faktor-faktor lain yang dapat memengaruhi sentimen publik, seperti faktor sosial, politik, dan budaya.

#### DAFTAR PUSTAKA

- [1] H. Fadilla, "Pengembangan Sektor Pariwisata untuk meningkatkan Pendapata Daerah Di Indonesia," vol. 2, no. 1, hal. 36–43, 2024, doi: 10.31080/BENEFIT.
- [2] V. R. Y. H. D dan Havidz Ageng Prakoso, "Penyelenggaraan Pagelaran Olahraga Balap Di Sirkuit Mandalika Sebagai Upaya Peningkatan Nation Branding Indonesia," *Moderat Jurnal Ilm. Ilmu Pemerintah.*, vol. 8, no. 2, hal. 284–302, 2022, doi: 10.25157/moderat.v8i2.2704.
- [3] A. Danartikanya, "5 Kendala Mandalika yang Wajib Diatasi demi MotoGP 2023 yang Lebih Lancar dan Meriah," 2022. <https://www.bola.net/otomotif/5-kendala-mandalika-yang-wajib-diatasi-demi-motogp-2023-yang-lebih-lancar-dan-meriah-4f422b.html>
- [4] H. B. Tambunan dan T. W. D. Hapsari, "Analisis Opini Pengguna Aplikasi New PLN Mobile

- 
- Menggunakan Text Mining,” *Petir*, vol. 15, no. 1, hal. 121–134, 2021, doi: 10.33322/petir.v15i1.1352.
- [5] S. Mujahidin, M. N. Hasyim, dan B. M. Pratama, “Implementasi Analisis Sentimen Opini Publik Mengenai Sirkuit Internasional Mandalika Pada Twitter Menggunakan Metode Multinomial Naïve Bayes Classifier,” *Bianglala Inform.*, vol. 10, no. 2, hal. 129–136, 2022, doi: 10.31294/bi.v10i2.13544.
- [6] F. F. Irfani, “Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine,” *JBMI (Jurnal Bisnis, Manajemen, dan Inform.)*, vol. 16, no. 3, hal. 258–266, 2020, doi: 10.26487/jbmi.v16i3.8607.
- [7] R. B. Afandi, T. F. Nurdiansyah, A. N. Ramadhani, dan A. P. Sari, “IMPLEMENTASI SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN APLIKASI ‘ MPStore - Super App UMKM ,”” hal. 565–570, 2024.
- [8] P. A. Sumitro, Rasiban, D. I. Mulyana, dan W. Saputro, “Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based,” *J-ICOM - J. Inform. dan Teknol. Komput.*, vol. 2, no. 2, hal. 50–56, 2021, doi: 10.33059/j-icom.v2i2.4009.
- [9] M. Haris, “Text Preprocessing,” hal. 7–19, 2009.
- [10] T. Mustaqim, *Sentiment Analysis Opini Pelantikan Kabinet Pemerintah Indonesia Tahun 2019 Menggunakan VADER dan Random Forest*. 2020.
- [11] Melita Ria, “Application of Term Frequency Inverse Document Frequency (TF-IDF) and Cosine Similarity Methods in Information Retrieval Systems to Know Web-Based Hadith Syarah (Case Study: Syarah Umdatil Ahkam),” *Tugas Akhir*, hal. 158, 2018.
- [12] DTRI WISUDAWATI, “Support Vector Machine,” hal. 14–41, 2020, [Daring]. Tersedia pada: <http://repository.unimus.ac.id>
- [13] A. M. Rizki dan A. L. Nurlaili, “Algoritme Particle Swarm Optimization ( PSO ) untuk Optimasi Perencanaan Produksi Agregat Multi-Site pada Industri Tekstil Rumahan,” hal. 1–9.
- [14] Z. I. Alfianti, “Analisis Sentimen Review Kosmetik Pada Website Femaledaily Menggunakan Metode Naive Bayes Dan Support Vector Machine Berbasis Particle Swarm Optimization,” *Repository.Bsi.Ac.Id*, 2019, [Daring]. Tersedia pada: <https://repository.bsi.ac.id/index.php/unduh/item/340509/Full-Tesis.pdf>
- [15] Abdul Muiz Khalimi, “Perhitungan Confusion Matrix Multi Class,” 2020. <https://www.pengalaman-edukasi.com/2020/11/menghitung-confusion-matrix-3-kelas.html>