# Analisis Sentimen Cuitan X terhadap Pendaki Asing Gunung Rinjani Menggunakan Algoritma SVM dan Random Forest

**Desy Candra Novitasari[1], Rahmi Rizkiana Putri[2]**
[1]Jurusan Bisnis Digital, Fakultas Ekonomi dan Bisnis, Universitas Kepanjen
[2]Jurusan Teknik Informatika, Fakultas Teknik Elektro dan Teknologi Informasi, Institut Teknologi Adhi Tama Surabaya
Email: [1]desycandra.045@gmail.com, [2]rahmi@itats.ac.id

***Abstract.*** *Social media platforms have become significant sources of public opinion regarding various issues, including tourism activities. This study analyzes public sentiment towards foreign climbers on Mount Rinjani through X (Twitter) social media posts using Support Vector Machine (SVM) and Random Forest algorithms. Data collection employed web scraping techniques with six relevant keywords, resulting in 4,777 unique tweets after cleaning and duplicate removal. The lexicon-based labeling approach revealed an imbalanced distribution with 63.58% negative sentiment, 18.8% positive sentiment, and 17.63% neutral sentiment. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) was applied, creating a balanced dataset of 9,111 samples. Text preprocessing included case folding, normalization, tokenization, stopword removal, and stemming using Indonesian language tools. Feature extraction utilized TF-IDF vectorization with parameters optimized for Indonesian text analysis. The dataset was split into 70% training, 15% validation, and 15% testing using stratified sampling. Evaluation results demonstrated that SVM achieved superior performance with 95.7% accuracy, 96% precision, 95.7% recall, and 95.7% F1-score, while Random Forest achieved 94.4% accuracy, 94.4% precision, 94.4% recall, and 94.4% F1-score. The dominance of negative sentiment indicates public concerns regarding foreign climbing activities that require stakeholder attention. This research contributes to sentiment analysis methodology for Indonesian social media text and provides practical insights for sustainable tourism management in Mount Rinjani National Park.*
***Keywords:*** *sentiment analysis, social media, Mount Rinjani, support vector machine, random forest*

***Abstrak.*** *Platform media sosial telah menjadi sumber opini publik yang signifikan terhadap berbagai isu, termasuk aktivitas pariwisata. Penelitian ini menganalisis sentimen masyarakat terhadap pendaki asing di Gunung Rinjani melalui cuitan media sosial X menggunakan algoritma Support Vector Machine (SVM) dan Random Forest. Pengumpulan data menggunakan teknik web scraping dengan enam kata kunci relevan, menghasilkan 4,777 cuitan unik setelah pembersihan dan penghapusan duplikat. Pendekatan pelabelan berbasis leksikon menunjukkan distribusi tidak seimbang dengan 63.58% sentimen negatif, 18.8% sentimen positif, dan 17.63% sentimen netral. Untuk mengatasi ketidakseimbangan kelas, diterapkan Synthetic Minority Oversampling Technique (SMOTE) yang menghasilkan dataset seimbang sebanyak 9,111 sampel. Preprocessing teks meliputi case folding, normalisasi, tokenisasi, stopword removal, dan stemming menggunakan tools bahasa Indonesia. Ekstraksi fitur menggunakan vektorisasi TF-IDF dengan parameter yang dioptimalkan untuk analisis teks Indonesia. Dataset dibagi menjadi 70% training, 15% validasi, dan 15% testing menggunakan stratified sampling. Hasil evaluasi menunjukkan SVM mencapai performa superior dengan akurasi 95.7%, presisi 96%, recall 95.7%, dan F1-score 95.7%, sementara Random Forest mencapai akurasi 94.4%, presisi 94.4%, recall 94.4%, dan F1-score 94.4%. Dominasi sentimen negatif mengindikasikan kekhawatiran publik terhadap aktivitas pendakian asing yang memerlukan perhatian pemangku kepentingan. Penelitian ini berkontribusi pada metodologi analisis sentimen teks media sosial Indonesia dan memberikan insight praktis untuk pengelolaan pariwisata berkelanjutan di Taman Nasional Gunung Rinjani.*
***Kata Kunci:*** *analisis sentimen, media sosial, Gunung Rinjani, support vector machine, random forest*

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*

## 1. Pendahuluan

Pada era saat ini, kebanyakan orang cenderung menggunakan perkembangan teknologi yang ada dalam melakukan segala hal dalam kehidupan mereka (Vindua & Zailani, 2023). Miliaran data telah diproduksi setiap harinya, sekitar 80% dari mereka dalam berbentuk teks (Ziegler et al., 2020). Pesatnya perkembangan teknologi informasi dan komunikasi telah merevolusi cara individu berinteraksi dan mengutarakan pandangan mereka, khususnya melalui media social (Manalu et al., 2022). Platform media social seperti X (sebelumnya Twitter) kini berfungsi sebagai kanal utama bagi pengguna untuk mendistribusikan pengalaman, menyampaikan kritik, dan menunjukkan sentimen mereka terkait topik atau isu-isu tertentu (Himawan et al., 2021).

Sentimen dapat diartikan sebagai proses pengungkapan atau suatu ekspresi dari berbagai pendapat, emosi, dan penilaian yang tertuang dalam bentuk teks (Dayani et al., 2024). Berbagai cara seseorang mengungkapkan sentimen tersebut, bisa melalui media sosial, forum online, maupun platform berita (Indri Wika Astuti et al., 2022). Analisis sentimen, atau *sentiment analysis*, adalah sebuah pendekatan dalam ilmu komputer untuk secara otomatis mengidentifikasi, mengekstrak, dan mengklasifikasi sentimen yang terkandung dalam teks (Idris et al., 2023). Analisis sentimen, yang merupakan cabang dari penelitian text mining, telah berkembang menjadi alat yang sangat penting untuk memahami opini publik dan mengekstraksi informasi berharga dari data tekstual yang tidak terstruktur (Purnamasari et al., 2023). Metode ini mengkategorikan teks berdasarkan sentimen emosional, seperti negatif, positif, netral, atau sedih (Kafiar & Supatman, 2024). Dengan demikian, analisis sentimen berfungsi sebagai alat penting bagi bisnis untuk membuat keputusan strategis (Sujadi, 2022), meningkatkan tingkat kepuasan pelanggan, dan mengidentifikasi tren atau isu-isu yang membutuhkan tindak lanjut (Nada Adela et al., 2024). Bidang ini berkembang pesat karena kemampuannya untuk mengolah data tekstual dalam jumlah besar (Anjani et al., 2022) dan memberikan wawasan berharga bagi berbagai pihak, mulai dari perusahaan yang ingin memantau reputasi merek, politisi yang ingin memahami opini pemilih (Dayani et al., 2024), hingga lembaga riset yang mengkaji pandangan publik terhadap suatu isu (Syahrohim et al., 2024).

Proses dalam melakukan analisis sentimen melalui penerapan metode pemrosesan bahasa alami (NLP) dan pembelajaran mesin atau biasa dikenal dengan *Machine Learning* (Manalu et al., 2022). Bagi peneliti dan praktisi dimungkinkan untuk mengidentifikasi, melakukan analisis mendalam, serta mengkategorikan pendapat, sentimen, dan penilaian yang terkandung dalam data tekstual (Indriyani et al., 2022). Banyak peneliti telah mengembangkan berbagai metode berbasis *machine learning* untuk mengolah data yang besar dalam bentuk teks. Beberapa metode yang populer dan sering digunakan antara lain *Support Vector Machine* (SVM), *Random Forest*, dan *Naive Bayes*. Setiap metode ini memiliki karakteristik matematis dan cara kerja yang berbeda dalam mempelajari pola dari data, yang pada akhirnya mempengaruhi akurasi dan performa model (Mursyid et al., 2024). Mengingat banyaknya pilihan metode yang tersedia dalam menganalisa sentimen dari teks pengguna, maka penting untuk melakukan perbandingan metode. Perbandingan metode analisis sentimen memungkinkan untuk mengidentifikasi algoritma yang paling sesuai dan memberikan performa terbaik untuk suatu kasus atau jenis data tertentu (Indriyani et al., 2022; Nada Adela et al., 2024). Hasil perbandingan ini menjadi dasar untuk memilih metode yang paling efisien dan akurat, sehingga analisis yang dilakukan bisa menghasilkan kesimpulan yang lebih valid dan dapat diandalkan (Dayani et al., 2024).

Untuk melakukan analisis sentimen yang efektif, pemilihan algoritma pembelajaran mesin yang tepat menjadi faktor krusial dalam menentukan akurasi dan reliabilitas hasil klasifikasi (Bustomy, 2021). *Support Vector Machine* (SVM) telah terbukti sebagai salah satu algoritma yang paling efektif untuk klasifikasi teks, dengan kemampuan untuk menangani data berdimensi tinggi dan menghasilkan akurasi yang tinggi dalam berbagai aplikasi analisis sentimen (Himawan et al., 2021). Penelitian sebelumnya menunjukkan bahwa SVM memiliki dasar teoritis yang kuat dan mampu melakukan

klasifikasi lebih akurat daripada kebanyakan algoritma lain dalam banyak aplikasi, dengan tingkat akurasi yang dapat mencapai hingga 98% dalam beberapa kasus (Adela et al., 2024).

Sementara itu, algoritma *Random Forest* juga telah menunjukkan performa yang excellent dalam klasifikasi sentimen, terutama dalam menangani dataset yang kompleks dan besar . Random Forest memiliki keunggulan dalam mengurangi overfitting dan memberikan hasil yang stabil, serta kemampuan untuk menangani missing values dan memberikan estimasi importance dari fitur-fitur yang digunakan (Saepudin et al., 2024). Beberapa penelitian telah menunjukkan bahwa Random Forest dapat mencapai akurasi hingga 94% dalam klasifikasi sentimen pada ulasan aplikasi e-commerce (Saepudin et al., 2024).

Proses analisis sentimen melibatkan beberapa tahapan penting, dimulai dari pengumpulan data, preprocessing teks, ekstraksi fitur, hingga klasifikasi menggunakan algoritma pembelajaran mesin. Tahapan preprocessing yang mencakup *case folding, tokenizing, stopword removal*, dan *stemming* menjadi sangat penting untuk mengoptimalkan kualitas data sebelum proses klasifikasi (Muzayyanah et al., 2024). Selain itu, penggunaan teknik pembobotan seperti TF-IDF (*Term Frequency-Inverse Document Frequency*) telah terbukti efektif dalam meningkatkan performa algoritma klasifikasi dengan memberikan bobot yang tepat pada kata-kata yang signifikan (Adrian et al., 2021).

Penelitian terdahulu dalam bidang analisis sentimen telah menunjukkan variasi hasil akurasi yang signifikan tergantung pada domain aplikasi, jenis data, dan algoritma yang digunakan. SVM dengan kernel linear dapat menghasilkan ketepatan klasifikasi yang lebih tinggi dibandingkan dengan kernel RBF dalam analisis sentimen vaksin COVID-19 (Hadianti et al., 2022). Sementara itu, penelitian oleh Allorerung dan Rismayani (2023) menunjukkan bahwa SVM mencapai akurasi 80% dalam analisis sentimen ulasan aplikasi streaming, sementara Naive Bayes Classifier mencapai akurasi yang lebih rendah (Palinggik Allorerung et al., 2023).

Optimasi algoritma juga menjadi aspek penting dalam meningkatkan performa analisis sentimen. Beberapa penelitian telah menerapkan teknik optimasi seperti *Particle Swarm Optimization* (PSO) dan *Genetic Algorithm* (GA) untuk meningkatkan akurasi algoritma SVM. Pada penelitian sebelumnya, kombinasi SVM dengan PSO dan GA dapat mencapai akurasi hingga 95% dalam analisis sentimen pemekaran Papua di Twitter (Purnamasari et al., 2023). Demikian pula, penelitian oleh Saepudin et al. (2024) menunjukkan bahwa SVM dengan feature selection PSO dapat meningkatkan akurasi dari 87.50% menjadi 89.50%.

Dalam konteks pariwisata dan media sosial, beberapa penelitian telah mengeksplorasi analisis sentimen untuk memahami persepsi publik terhadap berbagai isu. Namun, penelitian yang secara spesifik menganalisis sentimen cuitan media sosial terhadap pendaki asing di Gunung Rinjani masih sangat terbatas. Kebanyakan penelitian analisis sentimen dalam domain pariwisata masih terfokus pada analisis ulasan aplikasi atau *platform e-commerce*, sementara analisis sentimen terkait isu pariwisata alam dan dampaknya terhadap masyarakat lokal belum banyak dieksplorasi (Alghifari et al., 2022). Dalam penelitian ini, penulis fokus pada kasus Analisis Sentimen terhadap pendaki asing di Gunung Rinjani. Data yang digunakan bersumber dari komentar dan diskusi publik di media sosial X (Twitter) selama dua bulan terakhir, yang berkaitan dengan kejadian, kecelakaan, dan aktivitas pendakian di Gunung Rinjani. Kesenjangan penelitian ini menjadi semakin penting mengingat meningkatnya jumlah wisatawan asing yang mendaki Gunung Rinjani dan dampak sosial-ekonomi yang ditimbulkannya. Analisis sentimen dapat membantu mengidentifikasi persepsi masyarakat terhadap kehadiran pendaki asing, baik yang bersifat positif maupun negatif, sehingga dapat menjadi dasar untuk pengembangan kebijakan pariwisata yang lebih responsif dan berkelanjutan. Selain itu, perbandingan performa algoritma SVM dan Random Forest dalam konteks spesifik pada penelitian ini dapat memberikan kontribusi metodologis yang berharga bagi penelitian analisis sentimen di masa depan.

Penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan melakukan analisis sentimen cuitan media sosial X terhadap pendaki asing di Gunung Rinjani menggunakan algoritma SVM dan Random Forest. Melalui pendekatan komparatif, penelitian ini akan mengevaluasi performa kedua algoritma dalam mengklasifikasikan sentimen dan memberikan insight tentang persepsi publik terhadap aktivitas pendakian asing di salah satu gunung tertinggi di Indonesia. Hasil penelitian ini

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*

diharapkan dapat memberikan kontribusi teoritis dalam pengembangan metode analisis sentimen serta kontribusi praktis bagi pengembangan strategi pengelolaan pariwisata yang berkelanjutan di Gunung Rinjani.

## 2. Tinjauan Pustaka

### 2.1 Analisis Sentimen dalam Media Sosial

Analisis sentimen merupakan cabang dari penelitian *text mining* yang melakukan proses pengklasifikasian dokumen teks untuk mengekstraksi pendapat, emosi, dan evaluasi tertulis seseorang tentang topik tertentu menggunakan teknik pemrosesan bahasa alami (Idris et al., 2023). Dalam konteks media sosial, platform seperti X (Twitter) telah menjadi sumber data yang kaya karena memungkinkan pengguna mengekspresikan pendapat secara real-time dan spontan (Purnamasari et al., 2023).

Media sosial berfungsi tidak hanya sebagai platform komunikasi, tetapi juga sebagai barometer sentimen publik. Himawan dan Eliyani (2021) menjelaskan bahwa pemerintah memanfaatkan media sosial seperti Twitter sebagai kanal interaksi dengan masyarakat, dimana informasi hasil interaksi tersebut menjadi umpan balik untuk mengetahui opini masyarakat terhadap kebijakan publik. Tantangan utama dalam analisis sentimen media sosial adalah volume data yang besar, dengan ribuan posting yang muncul setiap hari, serta penggunaan bahasa informal, singkatan, dan emoticon (Vindua & Zailani, 2023).

Aplikasi analisis sentimen telah menunjukkan relevansinya dalam berbagai domain, mulai dari evaluasi produk hingga pemantauan opini politik. Syahrohim et al. (2024) mendemonstrasikan bagaimana analisis sentimen dapat memberikan *insight* tentang performa algoritma klasifikasi dalam konteks media sosial, yang penting untuk pengembangan metodologi yang lebih baik dalam pemrosesan data tekstual tidak terstruktur.

### 2.2 *Support Vector Machine* dalam Klasifikasi Sentimen

Support Vector Machine telah terbukti sebagai salah satu algoritma pembelajaran mesin yang paling efektif untuk klasifikasi teks dan analisis sentimen. Keunggulan SVM terletak pada kemampuannya menemukan hyperplane optimal yang memisahkan kelas-kelas data dengan margin maksimal. Indriyani et al. (2022) menunjukkan bahwa SVM mempunyai dasar teoritis yang kuat dan melakukan klasifikasi lebih akurat daripada algoritma lain, dengan tingkat akurasi mencapai 87.27% dalam analisis sentimen.

Dalam praktiknya, SVM bekerja dengan memetakan data teks ke ruang fitur berdimensi tinggi melalui kernel functions. Adela et al. (2024) menunjukkan bahwa SVM memiliki akurasi tertinggi 63% dalam analisis sentimen aplikasi Seabank, mengungguli *Gaussian Naïve Bayes* dengan nilai 30%. Pemilihan kernel yang tepat sangat mempengaruhi performa, dimana Muzayyanah et al. (2024) menemukan bahwa kernel Linear menghasilkan akurasi 95.46% dibandingkan kernel RBF dengan 94.15%.

Optimasi SVM melalui teknik advanced menunjukkan peningkatan performa yang signifikan. Purnamasari et al. (2023) melaporkan bahwa kombinasi SVM dengan *Particle Swarm Optimization* dan *Genetic Algorithm* mencapai akurasi 95.00% dengan AUC 0.912, menunjukkan potensi besar teknik optimasi dalam meningkatkan performa klasifikasi sentimen.

### 2.3 *Random Forest* untuk Analisis Sentimen

*Random Forest* merepresentasikan pendekatan ensemble learning yang menggabungkan multiple decision trees untuk menghasilkan prediksi yang lebih robust. Algoritma ini menawarkan keunggulan dalam menangani overfitting, memberikan estimasi *feature importance*, dan bekerja baik pada dataset dengan *noise*. Saepudin et al. (2024) menunjukkan bahwa *Random Forest* mencapai

akurasi tertinggi 94% dalam analisis sentimen ulasan Shopee, mengungguli SVM (91%) dan *Logistic Regression* (86%).

Mekanisme *Random Forest* melibatkan pembangunan multiple decision trees melalui bootstrap sampling, dimana setiap tree memberikan vote untuk klasifikasi akhir. Herjanto dan Carudin (2024) menunjukkan performa yang konsisten dengan akurasi 74%, presisi 75%, recall 74%, dan f1-score 74% pada dataset 5000 ulasan aplikasi SIREKAP.

Keunggulan unik *Random Forest* adalah kemampuan memberikan *feature importance scores*, yang valuable dalam memahami kata-kata atau fitur yang paling berkontribusi terhadap klasifikasi sentimen. Robustness terhadap noise dan outliers juga menjadikan algoritma ini cocok untuk analisis sentimen media sosial yang sering mengandung typos, slang, dan bahasa informal (Herjanto & Carudin, 2024).

### 2.4 Penelitian Terkait dalam Domain Pariwisata dan Isu Sosial

Penelitian analisis sentimen dalam domain pariwisata menunjukkan perkembangan signifikan, meskipun sebagian besar terfokus pada ulasan aplikasi travel dan *platform e-commerce*. Beberapa penelitian telah mengeksplorasi isu sosial dan budaya terkait pariwisata, seperti Kafiar dan Supatman (2024) yang menganalisis sentimen netizen terhadap isu pembabatan hutan adat Papua dengan SVM, mencapai akurasi 67%.

Metodologi penelitian menunjukkan variasi menarik dalam pendekatan analisis sentimen. Astuti et al. (2022) menggunakan teknologi big data dengan 2.507.349 data Twitter dan algoritma *K-Nearest Neighbor*, mencapai akurasi tertinggi 76.19%. Scale penelitian ini menunjukkan potensi big data analytics dalam analisis sentimen pariwisata.

Gap penelitian yang teridentifikasi adalah minimnya penelitian yang menganalisis sentimen masyarakat terhadap aktivitas wisatawan asing di destinasi pariwisata alam Indonesia. Kebanyakan penelitian existing terfokus pada analisis sentimen aplikasi atau layanan digital, sementara aspek sosial-budaya dan dampak pariwisata terhadap masyarakat lokal belum banyak dieksplorasi. Selain itu, perbandingan sistematis antara SVM dan *Random Forest* dalam konteks analisis sentimen pariwisata masih terbatas.

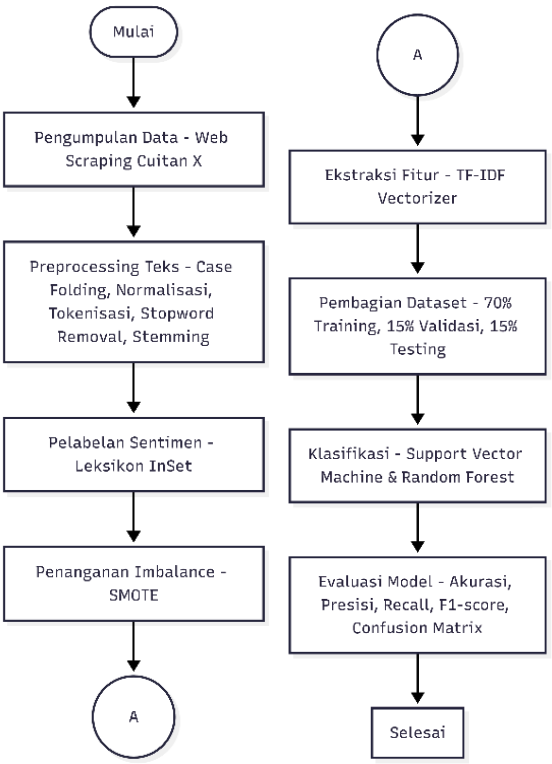
## 3. Metode Penelitian

### 3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimental untuk membandingkan performa algoritma Support Vector Machine (SVM) dan Random Forest dalam analisis sentimen cuitan media sosial X terhadap pendaki asing di Gunung Rinjani. Desain penelitian yang dipilih memungkinkan evaluasi sistematis terhadap kedua algoritma melalui serangkaian tahapan yang terstruktur, mulai dari pengumpulan data hingga evaluasi model.

Framework penelitian mengadopsi metodologi Knowledge Discovery in Database (KDD) yang terdiri dari lima tahapan utama: seleksi data, preprocessing, transformasi data, data mining, dan evaluasi. Pendekatan ini dipilih karena telah terbukti efektif dalam penelitian analisis sentimen dan memberikan struktur yang jelas untuk memastikan reprodusibilitas hasil penelitian. Berikut ini adalah algoritma dari metodologi penelitian ini dalam bentuk bagan atau *flowchart*:

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*




**Gambar 1 Diagram Alur Metodologi Penelitian**

## 3.2 Pengumpulan dan Persiapan Data

### 3.2.1 Sumber Data dan Strategi Pencarian

Data penelitian dikumpulkan dari platform media sosial X menggunakan teknik web scraping dengan kata kunci pencarian seperti Juliana Marins, Gunung Rinjani, Juliana Marins Rinjani, Kasus Rinjani, Pendaki Rinjani, Tewas Di Rinjani, dan Tragedi Rinjani. Pemilihan kata kunci ini didasarkan pada analisis pendahuluan terhadap trending topics dan diskusi publik terkait aktivitas pendakian asing di Gunung Rinjani. Proses pengumpulan dan pembersihan data menghasilkan distribusi dataset sebagaimana ditunjukkan pada Tabel 1. Tahap pembersihan ini penting untuk memastikan kualitas data dan menghindari bias yang dapat mempengaruhi performa model klasifikasi.

**Tabel 1 Distribusi Dataset Setelah Pembersihan**

| Tahapan | Jumlah Cuitan | Keterangan |
|---|---|---|
| Total data awal | 14185 | Data mentah hasil scraping |
| Data duplikat yang di hapus | 9393 | Data duplikat yang di eliminasi |
| Dataset final | 4777 | Data bersih untuk analisis |

### 3.2.2 Preprocessing Teks

Tahapan preprocessing teks dilakukan secara sistematis untuk mengoptimalkan kualitas data sebelum proses analisis sentimen. Proses preprocessing mengikuti rangkaian tahapan yang terstruktur sebagaimana ditunjukkan pada Tabel 2.

**Tabel 2 Tahapan Preprocessing Teks**

| Tahapan | Deskripsi | Tools/Library |
|---|---|---|
| Case Folding | Mengubah semua teks menjadi huruf kecil | Python string methods |
| URL Removal | Menghapus URL dan link | Regular Expression |
| HTML Cleaning | Menghilangkan HTML tags | Regular Expression |
| Emoji Removal | Menghapus emoji dan symbol | Emoji library |

| Tahapan | Deskripsi | Tools/Library |
|---|---|---|
| Symbol Removal | Menghilangkan tanda baca dan symbol | String punctuation |
| Normalisasi | Mengubah kata tidak baku menjadi baku | Kamus kata baku |
| Tokenisasi | Memecah teks menjadi token | NLTK tokenizer |
| Stopword Removal | Menghapus kata penghubung | NLTK stopwords |
| Stemming | Mengubah kata ke bentuk dasar | Sastrawi stemmer |

Normalisasi kata dilakukan menggunakan kamus kata baku bahasa Indonesia yang diperoleh dari dataset Kaggle "kamus-slag". Proses ini mengkonversi kata-kata tidak baku, singkatan, dan slang menjadi bentuk baku untuk meningkatkan konsistensi data. Tahap akhir preprocessing adalah stemming menggunakan library Sastrawi untuk mengubah setiap kata ke bentuk dasarnya.

### 3.3 Pelabelan Sentimen dan Penanganan Ketidakseimbangan Kelas

#### 3.3.1 Pelabelan Otomatis Menggunakan Leksikon

Pelabelan sentimen dilakukan menggunakan pendekatan berbasis leksikon dengan memanfaatkan dataset sentimen bahasa Indonesia (InSet) yang terdiri dari kamus kata positif dan negatif. Setiap cuitan dievaluasi berdasarkan jumlah kata positif dan negatif yang terkandung di dalamnya, dengan skor sentimen dihitung sebagai selisih antara jumlah kata positif dan negatif.

#### 3.3.2 Penanganan Ketidakseimbangan dengan SMOTE

Untuk mengatasi masalah ketidakseimbangan kelas yang dapat mempengaruhi performa model klasifikasi, penelitian ini menerapkan teknik Synthetic Minority Oversampling Technique (SMOTE). SMOTE bekerja dengan mensintesis sampel baru untuk kelas minoritas berdasarkan interpolasi antara sampel yang sudah ada, sehingga menciptakan dataset yang lebih seimbang tanpa duplikasi sederhana.

### 3.4 Ekstraksi Fitur dan Pembagian Dataset

#### 3.4.1 Vektorisasi TF-IDF

Ekstraksi fitur dilakukan menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) yang mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma machine learning. Konfigurasi parameter TF-IDF Vectorizer ditunjukkan pada Tabel 3.

**Tabel 3 Konfigurasi Parameter TF-IDF Vectorizer**

| Parameter | Nilai | Deskripsi |
|---|---|---|
| max_features | 10000 | Maksimal jumlah fitur yang digunakan |
| ngram_range | (1,2) | Menggunakan unigram dan bigram |
| min_df | 2 | Minimal muncul di 2 dokumen |
| max_df | 0.95 | Maksimal muncul di 95% dokumen |
| lowercase | True | Konversi ke huruf kecil |
| stop_words | None | Stopwords sudah dihapus sebelumnya |

#### 3.4.2 Stratified Sampling dan Pembagian Data

Dataset yang telah seimbang dibagi menjadi tiga subset menggunakan stratified sampling untuk mempertahankan proporsi kelas sentimen di setiap subset. Distribusi pembagian dataset ditunjukkan pada Tabel 4.

**Tabel 4 Pembagian Dataset**

| Subset | Jumlah Sampel | Persentase | Tujuan |
|---|---|---|---|
| Data Latih | 6377 | 70% | Training model |
| Data Validasi | 1367 | 15% | Validasi dan tuning |
| Data Uji | 1367 | 15% | Evaluasi final |
| Total | 9111 | 100% | |

### 3.5 Implementasi dan Evaluasi Model

#### 3.5.1 Konfigurasi Algoritma

Implementasi kedua algoritma klasifikasi menggunakan konfigurasi parameter yang dioptimalkan untuk analisis sentimen sebagaimana ditunjukkan pada Tabel 5.

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*

**Tabel 5 Konfigurasi Parameter Algoritma**

| Algoritma | Parameter | Nilai | Justifikasi |
|---|---|---|---|
| SVM | Kernel | 'poly' | Optimal untuk data teks |
| | random_state | 42 | Reproduksibilitas |
| | probability | True | Untuk probabilitas prediksi |
| Random Forest | n_estimators | 100 | Balance antara akurasi dan kecepatan |
| | random_state | 42 | Reproduksibilitas |

### 3.5.2 Metrik Evaluasi

Evaluasi model dilakukan menggunakan matrik evaluasi yang komprehensif sebagaimana ditunjukkan pada Tabel 6. Metrik-metrik ini dipilih karena memberikan gambaran lengkap tentang performa model dari berbagai aspek.

**Tabel 6 Metrik Evaluasi Model**

| Metrik | Persamaan | Deskripsi |
|---|---|---|
| Akurasi | $\frac{TP+TN}{TP+TN+FP+FN}$ | Proporsi prediksi yang benar secara keseluruhan |
| Presisi | $\frac{TP}{TP+FP}$ | Proporsi prediksi positif yang benar |
| Recall | $\frac{TP}{TP+FN}$ | Proporsi sampel positif yang berhasil diidentifikasi |
| F1-Score | $\frac{2 \times precision \times Recall}{precision+ Recall}$ | Harmonic mean antara presisi dan recall |

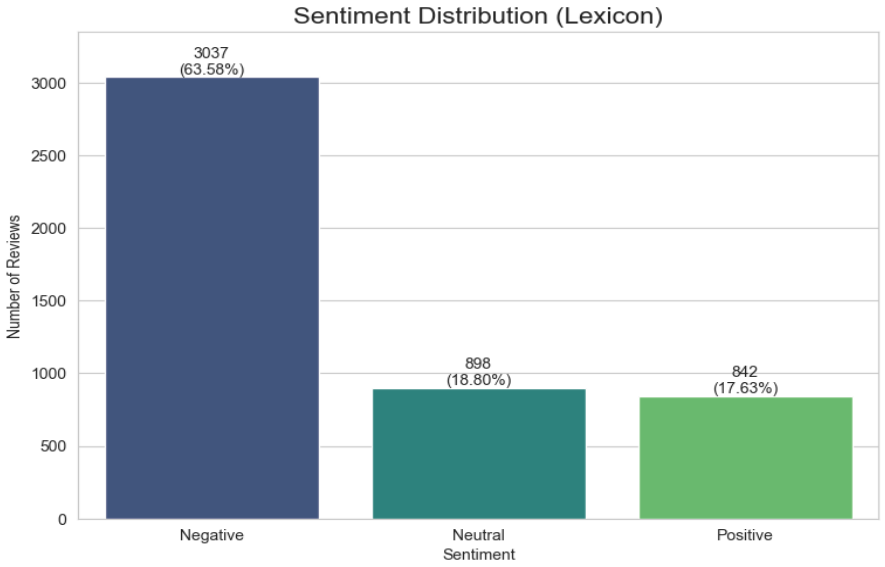
Keterangan: TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Selain metrik numerik, evaluasi juga mencakup confusion matrix untuk memberikan visualisasi detail tentang pola kesalahan klasifikasi dan classification report untuk analisis performa per kelas sentimen. Pendekatan evaluasi multi-metrik ini memastikan bahwa perbandingan antara SVM dan Random Forest dilakukan secara adil dan komprehensif.

## 4. Hasil dan Pembahasan

### 4.1 Karakteristik Dataset dan Distribusi Sentimen

Hasil pengumpulan data menggunakan enam kata kunci pencarian menghasilkan total 14.185 cuitan yang setelah melalui proses pembersihan dan penghapusan duplikat menjadi 4777 cuitan unik. Proses pelabelan sentimen menggunakan pendekatan berbasis leksikon menunjukkan distribusi yang tidak seimbang sebagaimana ditunjukkan pada Gambar 1.




**Gambar 2. Distribusi Sentimen**

Distribusi sentimen yang tidak seimbang ini mencerminkan kecenderungan yang lebih dominan dalam cuitan yang mengungkapkan kekhawatiran, kritik, atau respons negatif terkait aktivitas pendaki asing di Gunung Rinjani. Mayoritas cuitan negatif menunjukkan perhatian publik terhadap keselamatan dan dampak lingkungan.

## 4.2 Analisis Kata Kunci Berdasarkan Sentimen

Analisis distribusi kata yang paling sering muncul pada setiap kategori sentimen memberikan insight mendalam tentang topik dan isu yang mendominasi diskusi publik. Hanya kata kunci yang signifikan ditampilkan untuk menghindari pengulangan yang tidak perlu. Misalnya, kata 'kecewa' atau 'khawatir' mewakili sentimen negatif, sedangkan 'keindahan' dan 'pengalaman' mewakili sentimen positif. Tabel 7 menunjukkan 10 kata teratas untuk setiap kategori sentimen. Dan pada gambar 2. Menampilkan Word Cloud dari seluruh kelas.

**Tabel 7 Top 10 Kata Paling Sering Muncul Berdasarkan Sentimen**

| Ranking | Sentimen Positif | Frekuensi | Sentiment Netral | Frekuensi | Sentimen Negatif | Frekuensi |
|---|---|---|---|---|---|---|
| 1 | Rinjani | 852 | Rinjani | 876 | Rinjani | 3237 |
| 2 | Gunung | 705 | Gunung | 725 | Gunung | 2436 |
| 3 | Juliana | 374 | Juliana | 296 | Pendaki | 2111 |
| 4 | Marins | 300 | Pendaki | 267 | Korban | 962 |
| 5 | Jatuh | 198 | Marins | 241 | Juliana | 826 |
| 6 | Brasil | 194 | Jatuh | 219 | Jatuh | 776 |
| 7 | Selamat | 181 | Brasil | 167 | Tim | 659 |
| 8 | Pendaki | 170 | Evakuasi | 120 | Marins | 601 |
| 9 | Evakuasi | 132 | Korban | 108 | Sar | 573 |
| 10 | Korban | 116 | Selamat | 99 | Brasil | 556 |

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*


**Gambar 3. Wordcloud dari seluruh sentimen**

### 4.3 Penanganan Ketidakseimbangan Kelas dengan SMOTE

Penerapan SMOTE berhasil mengatasi masalah ketidakseimbangan kelas dalam dataset sebagaimana ditunjukkan pada Tabel 8.

**Tabel 8 Perbandingan Distribusi Dataset Sebelum dan Sesudah SMOTE**

| Sentimen | Sebelum SMOTE | | Sesudah SMOTE | |
|---|---|---|---|---|
| | Jumlah | Persentase | Jumlah | Persentase |
| Positif | 842 | 17.6% | 3037 | 33.3% |
| Netral | 898 | 18.8% | 3037 | 33.3% |
| Negatif | 3037 | 63.6% | 3037 | 33.3% |
| Total | 4777 | 100% | 9.111 | 100% |
| Imbalance Ratio | 3.61 | | 1.00 | |

### 4.4 Performa Model Klasifikasi

#### 4.4.1 Hasil Evaluasi Keseluruhan

Evaluasi performa kedua model pada dataset uji menunjukkan hasil yang sangat baik untuk kedua algoritma. SVM sedikit lebih unggul karena kemampuannya menangani data berdimensi tinggi dengan variasi kata yang kompleks, sedangkan Random Forest tetap stabil namun sedikit lebih rendah akurasinya. Tabel 4.5 menyajikan perbandingan metrik evaluasi keseluruhan.

**Tabel 4 Perbandingan Performa Model pada Dataset Uji**

| Model | Akurasi | Presisi | Recall | F1-Score |
|---|---|---|---|---|
| SVM | 95.70% | 96.00% | 95.70% | 95.70% |
| Random Forest | 94.40% | 94.40% | 94.40% | 94.40% |

#### 4.4.2 Analisis Performa Per Kelas Sentimen

Tabel 4.6 dan 4.7 menunjukkan breakdown performa masing-masing model untuk setiap kelas sentimen. Dan pada gambar 3 divisualisasikan confusion matrix pada setiap model

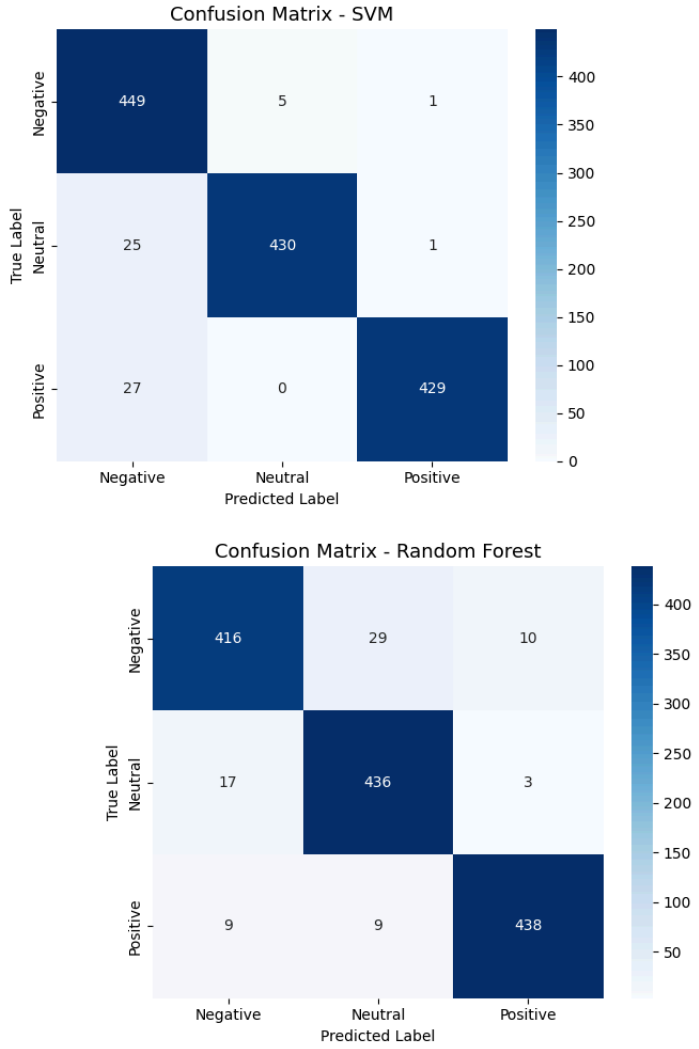
**Tabel 4 Performa Support Vector Machine Per Kelas**

| Sentimen | Presisi | Recall | F1-Score | Support |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Positif | 100% | 94% | 97% | 456 |
| Netral | 99% | 94% | 97% | 456 |
| Negatif | 90% | 99% | 94% | 456 |
| Rata-rata | 96% | 96% | 96% | 1367 |

**Tabel 4 Performa Random Forest Per Kelas**

| Sentimen | Presisi | Recall | F1-Score | Support |
|---|---|---|---|---|
| Positif | 94% | 91% | 97% | 456 |
| Netral | 92% | 96% | 94% | 456 |
| Negatif | 94% | 91% | 93% | 455 |
| Rata-rata | 94% | 94% | 94% | 1367 |




**Gambar 4 Confusion Matrix**

## 4.5 Analisis Kesalahan Klasifikasi

Analisis confusion matrix dari kedua model menunjukkan pola kesalahan yang serupa, dimana kesalahan klasifikasi paling sering terjadi antara sentimen negatif dan positif. Beberapa cuitan yang ambigu atau mengandung kata-kata ganda sering salah diklasifikasikan. Contoh: 'Pendakian di Rinjani sangat berbahaya dan menakutkan, harus lebih hati-hati!' diklasifikasikan positif padahal negatif, dan 'Walaupun pendakian di Rinjani sangat menantang, saya sangat menikmati petualangan ini!' diklasifikasikan negatif padahal positif. Sentimen netral menunjukkan tingkat akurasi tertinggi pada

*INTEGER: Journal of Information Technology, Vol 10, No 2, September 2025: 239-251*
*ISSN: 2579-566X (Online)*
*ISSN: 2477-5274 (Print)*

kedua model, mengindikasikan bahwa cuitan dengan karakteristik informatif lebih mudah diidentifikasi dibandingkan cuitan yang mengandung muatan emosional.

### 4.6 Implikasi Praktis dan Teoretis

Hasil penelitian ini memberikan kontribusi praktis yang signifikan untuk pemahaman sentimen publik terhadap aktivitas pendaki asing di Gunung Rinjani. Dominasi sentimen negatif dalam diskusi publik mengindikasikan adanya kekhawatiran yang perlu mendapat perhatian dari pemangku kepentingan. Dari perspektif metodologis, penelitian ini mendemonstrasikan efektivitas kedua algoritma dalam analisis sentimen bahasa Indonesia, dengan SVM menunjukkan sedikit keunggulan. Penerapan SMOTE yang berhasil juga memberikan insight penting tentang pentingnya preprocessing data dalam machine learning.

## 5. Kesimpulan dan Saran

### 5.1 Kesimpulan

Berdasarkan hasil penelitian analisis sentimen cuitan media sosial X terhadap pendaki asing di Gunung Rinjani menggunakan algoritma Support Vector Machine (SVM) dan Random Forest, dapat disimpulkan bahwa mayoritas cuitan mengandung sentimen negatif sebesar 63.58%, disusul sentimen positif sebesar 18.80% dan sentimen netral sebesar 17.63%. Hal ini menunjukkan bahwa masyarakat cenderung mengekspresikan kekhawatiran terkait aktivitas pendaki asing, terutama mengenai keselamatan dan dampak lingkungan. Perbandingan performa kedua algoritma menunjukkan bahwa SVM memiliki keunggulan marginal namun konsisten dibandingkan Random Forest. SVM mencapai akurasi 95.7% dengan presisi 96%, recall 95.7%, dan F1-score 95.7%, sedangkan Random Forest memiliki akurasi 94.4% dengan presisi 94.4%, recall 94.4%, dan F1-score 94.4%. Keunggulan SVM ini disebabkan oleh kemampuannya menangani data berdimensi tinggi dan variasi kata yang kompleks, sehingga menghasilkan klasifikasi yang lebih konsisten. Penerapan SMOTE terbukti efektif dalam mengatasi ketidakseimbangan kelas, mengubah rasio dari 3.61 menjadi 1,00, yang secara signifikan meningkatkan performa kedua model klasifikasi. Analisis kata kunci menunjukkan fokus diskusi publik terkait lokasi seperti Rinjani dan Gunung, aktivitas seperti mendaki, tokoh terkait seperti Juliana dan Marins, serta aspek keselamatan seperti evakuasi dan kecelakaan, memberikan insight tentang isu-isu utama yang mendominasi persepsi publik.

### 5.2 Saran

Untuk penelitian selanjutnya, disarankan agar eksplorasi algoritma deep learning seperti LSTM atau BERT dilakukan untuk menangani konteks teks yang lebih kompleks dan ambigu, serta memperluas dataset dari berbagai platform media sosial agar representasi opini publik lebih komprehensif. Implementasi analisis sentimen secara real-time dapat membantu pemantauan persepsi masyarakat secara kontinu, sementara penelitian longitudinal dapat memberikan wawasan tentang evolusi sentimen dari waktu ke waktu. Secara praktis, pengelola Taman Nasional Gunung Rinjani dapat memanfaatkan hasil penelitian ini untuk merancang strategi komunikasi yang lebih efektif dan mengurangi kekhawatiran masyarakat. Pemerintah daerah dapat menyesuaikan kebijakan pariwisata agar lebih responsif terhadap aspirasi masyarakat lokal, sedangkan industri pariwisata dapat menyesuaikan layanan untuk meningkatkan pengalaman wisata dan implementasi sustainable tourism. Secara metodologis, pengembangan corpus sentimen bahasa Indonesia yang lebih lengkap, eksplorasi teknik ensemble learning, penggunaan word embeddings untuk representasi fitur yang lebih semantik, serta pengembangan pipeline preprocessing teks bahasa Indonesia otomatis yang lebih robust menjadi saran penting untuk penelitian lanjutan agar analisis sentimen lebih akurat dan replikatif.

## Referensi

Adela, C. N., Karnila, S., Sutedi, S., & Agarina, M. (2024). Analisis Ulasan Pengguna Aplikasi Seabank Dengan Support Vector Machine Dan Naïve Bayes. *Jurnal Tekno Kompak*, *18*(2), 441.

https://doi.org/10.33365/jtk.v18i2.4156

Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati, N. A. (2021). Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB. *Jurnal Informatika Upgris*, *7*(1), undefined-undefined. https://doi.org/10.26877/JIU.V7I1.7099

Alghifari, D. R., Edi, M., & Firmansyah, L. (2022). Implementasi Bidirectional LSTM untuk Analisis Sentimen Terhadap Layanan Grab Indonesia. *Jurnal Manajemen Informatika (JAMIKA)*, *12*(2), 89–99. https://doi.org/10.34010/jamika.v12i2.7764

Anjani, A. M., Chamid, A. A., & Murti, A. C. (2022). Analisis Sentimen Kaum LGBT pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes. *Journal.Unisnu.Ac.Id*, *1*(2). https://journal.unisnu.ac.id/JTINFO/article/view/259

Bustomy, H. (2021). Analisa sentimen data text preprocessing pada data mining dengan menggunakan machine learning. *Journal.Ubm.Ac.Id*, *4*(2), 16–22. https://doi.org/10.30813/jbase.v4i2.3000

Dayani, A., KomtekInfo, G. N.-J., & 2024, U. (2024). Analisis Sentimen Terhadap Opini Publik pada Sosial Media Twitter Menggunakan Metode Support Vector Machine. *Jkomtekinfo.Org*. https://jkomtekinfo.org/ojs/index.php/komtekinfo/article/view/439

Hadianti, S., Yosep Tember, F., Mandiri, N., Raya, J., No, J., Melayu, C., & Timur, J. (2022). Analisis Sentimen COVID-19 di Twitter Menggunakan Metode Naïve Bayes dan SVM. *Jurnal Teknologi Informasi)*, *6*(1), 58–63.

Herjanto, M. F. Y., & Carudin, C. (2024). Analisis Sentimen Ulasan Pengguna Aplikasi Sirekap Pada Play Store Menggunakan Algoritma Random Forest Classifer. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2), 1204–1210. https://doi.org/10.23960/jitet.v12i2.4192

Himawan, R., Penelitian, E. E.-J. (Jurnal E. dan, & 2021, undefined. (2021). Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi. *Jurnal.Untan.Ac.IdRD Himawan, E EliyaniJEPIN (Jurnal Edukasi Dan Penelitian Informatika), 2021•jurnal.Untan.Ac.Id*. https://jurnal.untan.ac.id/index.php/jepin/article/view/41728

Idris, I., Mustofa, Y., And, I. S.-J. J. of E., & 2023, U. (2023). Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM). *Ejurnal.Ung.Ac.IdISK Idris, YA Mustofa, IA SalihiJambura Journal of Electrical and Electronics Engineering, 2023•ejurnal.Ung.Ac.Id*, *5*(1). https://doi.org/10.1177/0165551510388123

Indri Wika Astuti, N. N., Dwi Suarjaya, I. M. A., & Raharja, I. M. S. (2022). *ANALISIS SENTIMEN MASYARAKAT TERHADAP PELAKSANAAN LAYANAN KESEHATAN SELAMA MASA PANDEMI DI INDONESIA MENGGUNAKAN TEKNOLOGI BIG DATA Ni Nyoman Indri Wika Astuti a1 , I Made Agus Dwi Suarjaya a2 , I Made Sunia Raharja b3*. *3*(2).

Indriyani, E. R., Paradise, P., & Wibowo, M. (2022). Perbandingan Metode Naïve Bayes dan Support Vector Machine Untuk Analisis Sentimen Terhadap Vaksin Astrazeneca di Twitter. *Jurnal Media Informatika Budidarma*, *6*(3), 1545. https://doi.org/10.30865/mib.v6i3.4220

Kafiar, A. M., & Supatman, S. (2024). ANALISIS SENTIMEN NETIZEN TERHADAP ISU PEMBABATAN HUTAN ADAT PAPUA MELALUI TAGAR #ALLEYESONPAPUA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(4), 8129–8135. https://doi.org/10.36040/JATI.V8I4.10672

Manalu, D. R., L. Tobing, M. C., & Yohanna, M. (2022). Analisis Sentimen Twitter Terhadap Wacana Penundaan Pemilu Dengan Metode Support Vector Machine. *METHOMIKA Jurnal Manajemen Informatika Dan Komputerisasi Akuntansi*, *6*(6), 149–156. https://doi.org/10.46880/jmika.vol6no2.pp149-156

Mursyid, R., Computer, A. I.-J. of I. and, & 2024, undefined. (2024). Perbandingan Akurasi Metode Analisis Sentimen Untuk Evaluasi Opini Pengguna Pada Platform Media Sosial (Studi Kasus: Twitter). *Ejournal.Unesa.Ac.Id*, *06*. https://ejournal.unesa.ac.id/index.php/jinacs/article/download/61322/46931

Muzayyanah, A., … R. P.-I. I., & 2024, undefined. (2024). ANALISIS SENTIMEN PADA ULASAN APLIKASI EHADRAH DI GOOGLE PLAYSTORE MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM). *Jom.Fti.Budiluhur.Ac.IdAB Muzayyanah, RE Pawening, Z ArifinIDEALIS: InDonEsiA JournaL Information System, 2024•jom.Fti.Budiluhur.Ac.Id*, *7*(2), 258–266. https://jom.fti.budiluhur.ac.id/IDEALIS/article/view/3250

Nada Adela, C., Karnila, S., & Agarina, M. (2024). *Analisis Ulasan Pengguna Aplikasi Seabank Dengan Support Vector Machine Dan Naïve Bayes*. *18*(2), 441–453.

Palinggik Allorerung, P., Perintis Kemerdekaan, J. K., Tamalanrea, K., Makassar, K., & Selatan, S. (2023). Sentiment Analysis on WeTV App Reviews on Google Play Store Using NBC and SVM Algorithms. *Sistemasi.Ftik.Unisi.Ac.IdPP Allorerung, R RismayaniSISTEMASI, 2023•sistemasi.Ftik.Unisi.Ac.Id*, *12*(2), 2540–9719. https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/2518

Purnamasari, D., Anshary, M., AITI, R. R.-, & 2023, undefined. (2023). Particle Swarm Optimization dan Genetic Algorithm untuk analisis sentimen pemekaran Papua di Twitter berbasis Support Vector Machine. *Ejournal.Uksw.EduD Purnamasari, MAK Anshary, R RiantoAITI, 2023•ejournal.Uksw.Edu*, *20*(Agustus), 177–190. https://ejournal.uksw.edu/aiti/article/view/7747

Saepudin, A., Aryanti, R., Fitriani, E., Royadi, R., & Ardiansyah, D. (2024). Analisis Sentimen Pemanfaatan Artificial Intelligence di Dunia Pendidikan Menggunakan SVM Berbasis Particle Swarm Optimization. *Computer Science (CO-SCIENCE)*, *4*(1), 71–79. https://doi.org/10.31294/coscience.v4i1.2921

Sujadi, H. (2022). Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Metode Naive Bayes Classifier Dan Support Vector Machine. *INFOTECH Journal*, *8*(1), 22–27. https://doi.org/10.31949/infotech.v8i1.1883

Syahrohim, I., Saputra, S. D., Saputra, R. W., Pranatawijaya, V. H., & Priskila, R. (2024). PERBANDINGAN ANALISIS SENTIMEN SETELAH PILPRES 2024 DI TWITTER MENGGUNAKAN ALGORITMA MACHINE LEARNING. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2), 2830–7062. https://doi.org/10.23960/JITET.V12I2.4249

Vindua, R., & Zailani, A. U. (2023). Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python. *JURIKOM (Jurnal Riset Komputer)*, *10*(2), 479. https://doi.org/10.30865/JURIKOM.V10I2.5945

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Paul, D. A., Geoffrey, C., & Openai, I. (2020). Fine-tuning language models from human preferences. *Arxiv.Org*. https://arxiv.org/abs/1909.08593