Jurnal Ilmiah Teknik Informatika dan Komunikasi
Volume 6, Nomor 1, Maret 2026
E-ISSN : 2827-7945; P-ISSN : 2827-8127; Hal. 361-378
DOI: https://doi.org/10.55606/juitik.v6i1.2062
Tersedia: https://journal.sinov.id/index.php/juitik

Analisis Sentimen Pengunjung terhadap Tempat Wisata di Yogyakarta Menggunakan *Support Vector Machine (SVM)* dan *Linear Discriminant Analysis (LDA)*

Yuliana Putri[1*], Esi Putri Silmina[2]
[1-2] Program Studi Teknologi Informasi, Universitas Aisyiyah Yogyakarta, Indonesia
*Email: pyuliana314@gmail.com[1], esiputrisilmina@unisayogya.ac.id[2]

**Penulis Korespondensi:* pyuliana314@gmail.com

Abstract *Tourism is a strategic sector that plays an important role in supporting regional economic development, including in the Special Region of Yogyakarta as one of Indonesia's leading tourist destinations. High tourism activity has led to an increasing number of visitor reviews on online platforms such as TripAdvisor, which can be utilized to understand tourists' perceptions and satisfaction levels. The large volume of reviews makes manual analysis ineffective, thus requiring an automated, computation-based approach. This study aims to analyze the sentiment of tourism reviews in Yogyakarta City using the Support Vector Machine (SVM) algorithm with the addition of the Linear Discriminant Analysis (LDA) dimensionality reduction method. The dataset consists of 5,000 tourism reviews that have undergone preprocessing stages, including data cleaning, case folding, tokenization, normalization, stopword removal, and stemming. Text feature representation was performed using the TF-IDF method. Model evaluation was conducted using an 80:20 split between training and testing data, with performance measured using accuracy, precision, recall, and F1-score metrics. The results indicate that the application of LDA to the SVM model improves the balance of classification performance, particularly in precision (77.71%) and F1-score (79.5%), despite a slight decrease in accuracy. Sentiment classification results are dominated by positive sentiment (92.4%), reflecting generally favorable tourist perceptions of destinations in Yogyakarta. This study is expected to serve as a reference for tourism managers in evaluating service quality and facilities based on visitor opinions.*

Keywords: *Linear-Discriminant Analysis; Sentiment Analysis; Support Vector Machine; Tourism; TripAdvisor.*

Abstrak. Pariwisata merupakan sektor strategis yang berperan penting dalam mendukung perekonomian daerah, termasuk di Daerah Istimewa Yogyakarta sebagai salah satu destinasi wisata unggulan di Indonesia. Tingginya aktivitas wisata mendorong meningkatnya jumlah ulasan pengunjung pada platform daring seperti TripAdvisor, yang berpotensi dimanfaatkan untuk memahami persepsi dan tingkat kepuasan wisatawan. Besarnya volume ulasan menyebabkan analisis manual menjadi tidak efektif sehingga diperlukan pendekatan otomatis berbasis komputasi. Penelitian ini bertujuan untu menganalisis sentimen ulasan wisata di Kota Yogyakarta menggunakan algoritma Support Vector Machine (SVM) dengan penambahan metode reduksi dimensi Linear Discriminant Analysis (LDA). Dataset yang digunakan berjumlah 5.000 ulasan wisata yang telah melalui proses preprocessing meliputi data cleaning, case folding, tokenisasi, normalisasi, stopword removal, dan stemming. Representasi fitur teks dilakukan menggunakan metode TF-IDF. Evaluasi model dilakukan menggunakan pembagian data latih dan data uji dengan rasio 80:20 serta metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa penerapan LDA pada model SVM mampu meningkatkan keseimbangan performa klasifikasi, khususnya pada nilai precision (77,71%) dan F1-score (79,5%) meskipun terjadi sedikit penurunan akurasi. Hasil klasifikasi sentimen didominasi oleh sentimen positif (92,4%), yang mencerminkan persepsi wisatawan yang cenderung baik terhadap destinasi wisata di Yogyakarta. Penelitian ini diharapkan dapat menjadi referensi bagi pengelola wisata dalam mengevaluasi kualitas layanan dan fasilitas berdasarkan opini pengunjung.

Kata kunci: Analisis Sentimen; Linear-Discriminant Analysis; Pariwisata; Support Vector Machine; TripAdvisor.

1. LATAR BELAKANG

Pariwisata merupakan salah satu sektor strategis yang berperan penting dalam meningkatkan perekonomian daerah. Daerah Istimewa Yogyakarta (DIY) dikenal sebagai destinasi wisata unggulan di Indonesia yang memiliki kekayaan alam, budaya, sejarah dan keramahan penduduknya. Tingginya minat wisatawan terhadap Yogyakarta mendorong

Artikel Masuk: 28 November 2025; Revisi: 27 Desember 2025; Diterima: 26 Januari 2026; Terbit: 29 Januari 2026

meningkatnya aktivitas digital pengunjung, khususnya dalam bentuk ulasan dan penilaian pada platform daring seperti TripAdvisor dan platform sejenisnya. Ulasan digital tersebut menggambarkan pengalaman, tingkat kepuasan serta persepsi pengunjung terhadap suatu destinasi yang dapat dimanfaatkan sebagai sumber informasi penting bagi pengelola wisata untuk meningkatkan kualitas layanan dan daya tarik tempat wisata (Ipmawati et al., 2024)

Seiring meningkatnya volume ulasan yang dihasilkan, analisis secara manual menjadi tidak efektif karena memerlukan waktu dan tenaga yang besar. Diperlukan pendekatan otomatis berbasis komputasi untuk mengolah dan menganalisis data tekstual dalam jumlah besar. Teknik yang dapat diimplementasikan adalah analisis sentimen, yaitu teknik untuk mengidentifikasi polaritas opini pengguna terhadap suatu objek ke dalam kategori sentimen positif, negatif atau netral (Aryanto & Mardhiyyah, 2024)

Ulasan pengunjung melalui TripAdvisor digunakan sebagai data yang dianalisis dengan tujuan memberi jawaban para pengunjung terhadap tempat wisata di Yogyakarta. Kemudian ulasan diproses untuk menghasilkan klasifikasi sentimen yang dapat digunakan para pengelola tempat wisata sebagai referensi tambahan untuk meningkatkan kualitas pelayanan, fasilitas atau infrastruktur. (Larasati, 2024)

Penelitian menggunakan SVM untuk menganalisis sentiment opini pengunjung terhadap objek wisata berdasarkan ulasan di Google Maps dan menunjukkan bahwa metode tersebut mencapai akurasi rata-rata sebesar 83,8% dalam mengklasifikasikan tingkat sentimen ulasan (positif, negatif, netral) membuktikan efektivitas SVM pada lingkup pariwisata digital (Ipmawati et al., 2024). Penelitian serupa dilakukan oleh (Syahlan et al., 2023) pada objek wisata Air Mancur Sri Baduga di Purwakarta, di mana SVM menghasilkan akurasi sebesar 81%, nilai precision mencapai 94% dan recall mencapai 99%. Menunjukkan bahwa algoritma SVM dapat memberikan performa yang cukup baik dalam penge-lompokan sentimen terhadap ulasan wisata. Berdasarkan beberapa penelitian yang telah dilakukan, metode SVM memberikan hasil evaluasi efektif pada kasus klasifikasi sentimen dengan topik yang berbeda. Walaupun Support Vector Machine (SVM) efektif dalam menangani data berdimensi tinggi, sejumlah studi terbaru menunjukkan bahwa biaya komputasi pelatihan SVM meningkat secara signifikan ketika jumlah fitur dan sampel besar, terutama pada data teks atau data high-dimensional. Setelah dilakukan analisis mendalam disebutkan bahwa komputasi kernel SVM dapat menjadi hambatan praktis dalam dataset besar, sehingga sering memerlukan teknik seleksi atau reduksi fitur untuk mengurangi kompleksitas perhitungan dan konsumsi memori tinggi (Almaspoor et al., 2021). Survey lain juga menunjukkan keterbatasan komputasi pada SVM karena sensitivitas terhadap parameter dan struktur data, terutama pada dataset yang

sangat besar dan berdimensi tinggi (Guido et al., 2024). Penulis memberikan tambahan dengan menambahkan metode Dimensionality Reduction (Winarnie et al., 2023) bahwa reduksi dimensi dilakukan untuk menyederhanakan struktur data namun tetap mempertahankan kemam-puan pemisahan antar kelas.

Tujuan penelitian ini untuk menganalisis sentimen yang berdasar dari ulasan pengunjung tempat wisata di Yogyakarta guna mengetahui persepsi mereka terhadap objek wisata tersebut, serta peningkatan teknik klasifikasi sentimen menggunakan metode SVM dengan LDA sebagai tambahan opini berbasis data bagi pengelola tempat wisata.

2. KAJIAN TEORITIS

Analisis Sentimen Dalam Pariwisata Digital

Perkembangan teknologi informasi telah mendorong meningkatnya pemanfaatan data ulasan daring sebagai sumber informasi strategis dalam sektor pariwisata. Platform seperti TripAdvisor dan Google Maps memungkinkan wisatawan untuk mengekspresikan pengalaman, kepuasan, serta persepsi mereka terhadap suatu destinasi wisata secara terbuka. Ulasan tersebut menjadi data tekstual yang bernilai tinggi, namun sulit dianalisis secara manual ketika jumlahnya besar. Oleh karena itu, analisis sentimen menjadi pendekatan yang relevan untuk mengklasifikasikan opini wisatawan ke dalam sentimen positif, negatif, maupun netral secara (Aryanto & Mardhiyyah, 2024).

Analisis sentimen merupakan bagian dari text mining yang bertujuan mengidentifikasi polaritas emosi atau opini dalam teks. Dalam konteks pariwisata, analisis sentimen dapat membantu pengelola destinasi memahami tingkat kepuasan pengunjung, mengidentifikasi kelemahan layanan, serta merumuskan strategi peningkatan kualitas fasilitas dan pelayanan berbasis data ulasan pengguna (Ipmawati et al., 2024)

Support Vector Machine (SVM) Dalam Klasifikasi Sentimen

Support Vector Machine (SVM) merupakan salah satu algoritma supervised learning yang banyak digunakan dalam klasifikasi teks karena kemampuannya menangani data berdimensi tinggi serta menghasilkan generalisasi model yang baik. SVM bekerja dengan mencari hyperplane optimal yang memaksimalkan jarak antar kelas, sehingga efektif digunakan pada data teks yang direpresentasikan dalam bentuk vektor TF-IDF (Chetna, 2025)

Berbagai penelitian menunjukkan bahwa SVM memiliki performa yang baik dalam analisis sentimen pada domain pariwisata. Ipmawati et al. (2024) melaporkan bahwa SVM mampu mencapai akurasi rata-rata sebesar 83,8% dalam mengklasifikasikan sentimen ulasan wisata berbasis Google Maps rangkum. Penelitian lain oleh (Syahlan et al., 2023) menunjukkan

performa SVM yang sangat tinggi dengan nilai precision mencapai 94% dan recall 99% pada klasifikasi sentimen ulasan objek wisata Air Mancur Sri Baduga. Hasil-hasil tersebut mengindikasikan bahwa SVM merupakan algoritma yang efektif untuk analisis sentimen berbasis ulasan wisata.

Permasalahan Dimensi Tinggi Pada Data Teks

Meskipun SVM memiliki keunggulan dalam menangani data berdimensi tinggi, beberapa penelitian menyebutkan adanya keterbatasan pada aspek komputasi ketika jumlah fitur dan ukuran dataset meningkat secara signifikan. Representasi teks menggunakan TF-IDF umumnya menghasilkan ribuan hingga puluhan ribu fitur, yang berdampak pada meningkatnya waktu pelatihan dan konsumsi memori (Almaspoor et al., 2021).

(Guido et al., 2024) Performa SVM sangat dipengaruhi oleh struktur data dan parameter model pada data berdimensi tinggi, sehingga diperlukan teknik seleksi fitur atau reduksi dimensi untuk mengendalikan kompleksitas tanpa menurunkan performa klasifikasi.

Linear Discriminant Analysis (LDA) Sebagai Reduksi Dimensi

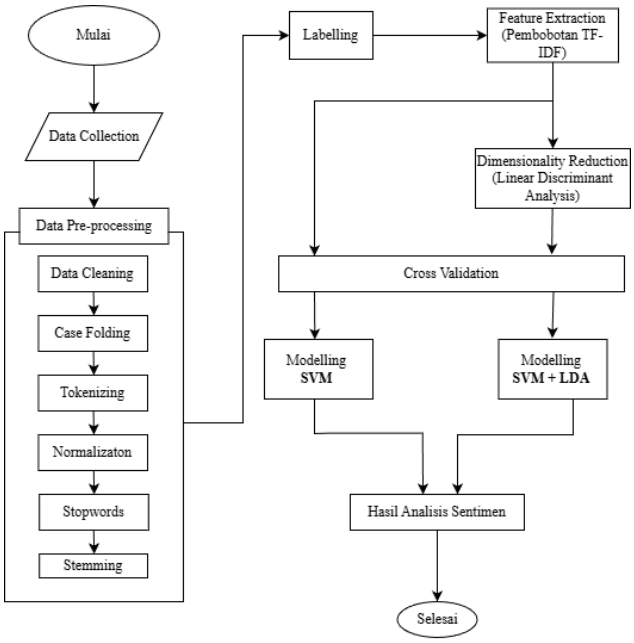
Linear Discriminant Analysis (LDA) adalah metode reduksi dimensi berbasis supervised learning yang memanfaatkan label kelas untuk mempertahankan separabilitas antar kelas, sehingga cocok digunakan dalam tugas klasifikasi sentimen. (Winarnie et al., 2023) Dalam penelitian text mining, LDA digunakan untuk menyederhanakan representasi fitur TF-IDF agar proses klasifikasi menjadi lebih efisien. (Lestari & Hutagalung, 2025) menyatakan bahwa kombinasi TF-IDF dengan reduksi dimensi mampu meningkatkan efisiensi komputasi serta menjaga stabilitas performa model klasifikasi. (Hanafi et al, 2025)juga membuktikan bahwa penerapan LDA sebelum SVM pada analisis sentimen ulasan wisata dapat meningkatkan keseimbangan performa klasifikasi antar kelas sentimen.

Kombinasi SVM Dan LDA Dalam Analisis Sentimen

Beberapa penelitian menunjukkan bahwa kombinasi SVM dan LDA efektif dalam mengatasi ketidakseimbangan kelas serta kompleksitas fitur pada data teks dengan meningkatkan precision dan F1-score, meskipun berpotensi menurunkan akurasi secara keseluruhan. (Hanafi et al, 2025) Penelitian ini menerapkan kombinasi SVM dan LDA untuk analisis sentimen ulasan wisata di Kota Yogyakarta guna menghasilkan klasifikasi yang lebih seimbang, efisien, dan representatif terhadap persepsi wisatawan.

3. METODE PENELITIAN

Metode penelitian ini menerapkan kuantitatif deskriptif dimana data dikumpulkan kemudian dianalisis sentimen ulasan atau review pengunjung. Data yang digunakan berfokus pada data ulasan tempat wisata di Yogyakarta yang diperoleh dari TripAdvisor dengan lebih dari 11.000 ulasan. Proses penelitian diawali dengan pengumpulan data, data pre-processing, labelling (menentukan label), feature extraction, dimensionality reduction dan evaluasi model.



Gambar 1. Tahapan Penelitian.

Penelitian ini dilakukan melalui tahapan pengumpulan data ulasan TripAdvisor, preprocessing teks, pelabelan sentimen, ekstraksi fitur, reduksi dimensi, dan evaluasi model analisis sentimen sesuai alur penelitian pada Gambar 1.

Data Collection

Data penelitian diperoleh dari instansi tempat magang dalam bentuk file Excel berisi 11.376 ulasan wisata dari TripAdvisor, yang kemudian diolah menggunakan Google Colab berbasis Python melalui tahap pembersihan data untuk memastikan konsistensi dan kesiapan analisis.*Data Pre-processing*

Pada proses ini dilakukan *preprocessing* data terlebih dahulu agar sesuai bentuk ejaan sehingga dapat di proses dengan mudah oleh *text minning*. *Text mining* merupakan merupakan

proses menganalisis teks untuk mengekstraksi informasi penting dan memperoleh hasil yang spesifik. Tahapan *preprocessing* data ditunjukan pada gambar

a. *Cleaning*

   Pada tahap ini dilakukan pembersihan terhadap data dengan menghapus emoji dan karakter aneh, angka, tanda baca serta *double space*. Atribut yang dihilangkan atau dihapus memang tidak berkaitan dengan tahap proses pengolahan data.

b. *Case Folding*

   Pada tahapan *case folding* seluruh teks akan diubah menjadi bentuk standar huruf kecil *(lowercase)*. Perubahan teks dilakukan agar data diproses dalam kondisi sama.

c. *Tokenizing*

   Proses selanjutnya, dilakukan *tokenizing* yaitu pemisahan kata dari setiap kalimat. Setiap kata dipisah berdasarkan *whitespace* atau setiap spasi dan menjadi token. Proses ini dibantu dengan pustaka *Natural Language Toolkit* (NLTK) dari *Python.*

d. *Normalizaation*

   Proses *normalization* atau normalisasi kata digunakan untuk mengubah kata yang tidak baku menjadi kata standar. Perubahan kata dipengaruhi oleh kamus *slangwords* yang digunakan. *Kaggle* menjadi sumber baru dalam mencari kamus untuk proses sentiment, sehingga lebih mudah dalam menyeleksi kamus yang cocok untuk penelitian ini (Diandra, 2022).

e. *Remove Stopwords*

   Proses selanjutnya adalah *remove stopwords.* Kata-kata yang tidak berhubungan dengan kasus akan dihilangkan. Proses ini dibantu dengan pustaka NLTK. Contoh kata yang tidak memiliki sentimen dan telah dihapus pada tahap pengolahan ini seperti kata *"agar", "harusnya", "hari", "di", "dan",* serta *"paling"*.

f. *Stemming*

   Sebelum proses *stemming* dilakukan, data terlebih dahulu diperiksa agar tidak terdapat nilai kosong atau *null* pada kolom teks ulasan. Proses *stemming* yaitu proses mengubah suatu kata bentukan menjadi kata dasar (Zulfikar, 2017). Setelah itu, proses stemming dilakukan menggunakan pendekatan berbasis aturan atau *rule-based* dengan menghapus akhiran tertentu tanpa menghilangkan angka dan kata penting agar makna teks tetap terjaga

Labelling

Data yang telah melewati pre-processing, kemudian diberikan label untuk menentukan jenis data berlabel positif, negatif atau netral. Proses labelling sentimen dilakukan

menggunakan kolom rating numerik yang tersedia dalam dataset. Pendekatan ini memanfaatkan penilaian pengguna berupa skor rating 1-5 sebagai acuan kepuasan pengunjung terhadap objek wisata. Penelitian sebelumnya (Ain et al., 2024) memanfaatkan rating sebagai dasar pelabelan sentimen. Penelitian tersebut menunjukkan bahwa rating dapat digunakan untuk memetakan sentimen pengguna terhadap tempat wisata secara efektif serta memberi informasi bagi pelaku industri pariwisata dalam mengevaluasi kualitas layanan.

Aturan pelabelan sentimen sebagai berikut:

a. Rating 4-5 dikategorikan sebagai sentimen positif
b. Rating 3 dikategorikan sebagai sentimen netral
c. Rating 1-2 dikategorikan sebagai sentimen negatif

$$\text{Sentimen} = \begin{cases} \text{Positif, jika rating} \geq 4 \\ \text{Netral, jika rating} = 3 \\ \text{Negatif, jika rating} \leq 2 \end{cases}$$

Label sentimen yang dihasilkan selanjutnya digunakan sebagai variabel target dalam proses pelatihan dan pengujian model klasifikasi *Support Vector Machine* (SVM) dengan reduksi dimensi menggunakan *Linear Discriminant Analysis* (LDA).

Feature Extraction

Feature extraction menggunakan TF-IDF dilakukan untuk mengubah teks ulasan menjadi representasi numerik berdasarkan frekuensi kemunculan kata. (Tri Putra et al., 2021). Tahapan pembobotan TF-IDF dalam penelitian ini secara sistematis sebagai berikut

Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode untuk mengkonversi data teks ke dalam format numerik dan memberikan bobot pada tiap kata atau fitur. TF mengukur frekuensi kemunculan kata dalam tiap dokumen sementara IDF menghitung jumlah dokumen yang memuat kata tersebut. Tujuan utama TF-IDF adalah mengidentifikasi kata kunci dalam dokumen (Wahyudi et al., 2024). Secara sistematis *Term Frequency* dapat dihitung menggunakan rumus sebagai berikut:

$$TF(t,d) = \frac{Jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{Jumlah\ total\ kata\ dalam\ dokumen\ d}$$

Frekuensi jumlah kata yang muncul dalam sebuah dokumen disebut *Term Frequency* (TF). Terkait rumus IDF yaitu:

$$IDF(d) = \frac{\text{Number of document}}{\text{Number of document with term t}}$$

Dimana N adalah jumlah dokumen dalam kumpulan dokumen dan n adalah jumlah dokumen yang mengandung kata tersebut. Setelah mendapatkan TF dan IDF, kemudian dapat

menghitung nilai TF-IDF yang merupakan hasil perkalian dari TF dan IDF. Rumus yang digunakan sebagai berikut:

$$F\text{-}IDF(t,d)=TF(t,d)\times IDF(d)$$

Keterangan :

t = Kata kunci term

d = Dokumen

t.d = Nilai TF-IDF untuk kata t dalam dokumen d

Tf = Banyaknya t (kata) yang dicari dalam dokumen

Idf = Banyak t kebalikan dari kata yang dicari

Dimensionality Reduction (Linear Discriminant Analysis)

Setelah ekstraksi fitur, reduksi dimensi menggunakan LDA diterapkan untuk menyederhanakan vektor TF-IDF guna meningkatkan efisiensi komputasi tanpa menghilangkan informasi penting antar kelas sentimen. (Lestari & Hutagalung, 2025). Pada penelitian ini memiliki 3 kelas. Setelah teks diubah menjadi vektor TF-IDF, LDA bergerak mengurangi jumlah fitur dari banyak sekali menjadi hanya 2 fitur saja *(n_components=2)* membuat data lebih sederhana tapi tetap bisa membedakan kelas sentimen (Hanafi et al, 2025).

Cross Validation

Data dibagi dengan rasio 80:20 menjadi data latih dan data uji untuk mengevaluasi kinerja SVM tanpa LDA dan SVM dengan LDA sebagai reduksi dimensi sebelum klasifikasi. (Zhahrina et al., 2025a), sehingga model belajar dari sebagian besar data dan diuji menggunakan data yang belum pernah digunakan sebelumnya. Hasil pembagian data pada tahap cross *validation* selanjutnya digunakan untuk membangun dua pemodelan.

Modelling

Pemodelan pertama, penelitian ini menggunakan algoritma *Support Vector Machine* (SVM) untuk melakukan klasifikasi sentimen ulasan wisata menjadi kelas positif, netral dan negatif. Model SVM menggunakan *kernel linear* karena sesuai untuk data teks berdimensi tinggi dan efektif dalam memisah antar kelas sentimen (Chetna, 2025).

Pada tahap pelatihan, SVM mempelajari hubungan antara fitur LDA dan label sentimen, kemudian model yang telah dilatih diuji untuk menilai kinerjanya.

Modelling SVM + LDA

Pemodelan kombinasi SVM dan LDA dilakukan dengan mereduksi fitur TF-IDF menggunakan LDA pada data latih dan menerapkannya pada data uji, kemudian hasil reduksi digunakan sebagai input SVM ber-kernel linear untuk pelatihan, prediksi, dan evaluasi kinerja klasifikasi sentimen. (Hanafi et al, 2025).

Model

Setelah pelatihan, kinerja model SVM dievaluasi menggunakan data uji hasil pembagian 80:20 untuk menilai kemampuan klasifikasi secara valid dan mengurangi risiko overfitting. (Zhahrina et al., 2025b).Model ini menggunakan TF-IDF dengan 800 fitur sebagai representasi teks, kemudian dilakukan reduksi dimensi menggunakan Linear Discriminant Analysis (LDA) menjadi 2 komponen untuk mengurangi dimensi dan mempertahankan informasi penting (Hanafi et al, 2025). Hasil reduksi LDA digunakan sebagai input SVM ber-kernel linear yang dilatih dan diuji, dengan evaluasi kinerja dilakukan menggunakan akurasi, precision, recall, F1-score, dan confusion matrix untuk menilai klasifikasi sentiment (Chetna, 2025). Hasil evaluasi menunjukkan kemampuan SVM+LDA dalam menangani data berdimensi tinggi dan memprediksi sentimen ulasan wisata secara efektif (Hanafi et al, 2025)

4. HASIL DAN PEMBAHASAN

Pengumpulan Data dan Labelling Sentimen

Data yang digunakan dalam penelitian ini berasal dari ulasan pengunjung tempat wisata di Yogyakarta yang dikumpulkan dari platform TripAdvisor. Ulasan tersebut merepresentasikan opini dan pengalaman wisatawan terhadap berbagai destinasi wisata yang ada di Yogyakarta. Dari keseluruhan data yang tersedia, dilakukan proses pembatasan jumlah data (sampling) guna meningkatkan efisiensi pemrosesan dan analisis, khususnya dalam penerapan metode machine learning. Setelah melalui tahap sampling, diperoleh sebanyak 5.000 dataset ulasan yang selanjutnya digunakan dalam proses analisis sentimen. Jumlah dataset tersebut dinilai cukup representatif untuk menggambarkan kecenderungan sentimen wisatawan sekaligus menjaga keseimbangan antara akurasi model dan efisiensi komputasi

Kode Program

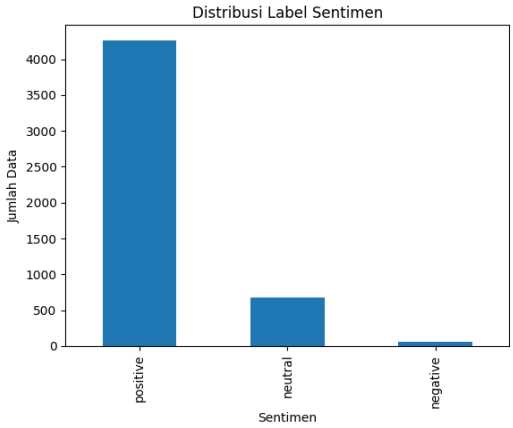
```
print("\n=== DATA COLLECTION ===")
DATA_PATH='/content/drive/MyDrive/dataset/attractions-reviews-sentiment.csv'
data = pd.read_csv(DATA_PATH, sep=';')
MAX_DATA = 5000
if len(data) > MAX_DATA:
   data = data.sample(MAX_DATA, random_state=42)
print("Jumlah data digunakan:", len(data))
```

Metode *random sampling* menggunakan *random_state=42* agar hasil eksperimen bersifat konsisten dan dapat direproduksi. Gambar 3 dan 4 menunjukkan jumlah data yang digunakan dalam penelitian dan distribusi label sentiment pada data.

```
=== DATA COLLECTION ===
Jumlah data digunakan: 5000
sentiment
positive    4263
neutral      675
negative      62
Name: count, dtype: int64
```

Gambar 3. Jumlah data yang digunakan.

Distribusi Label Sentimen



Gambar 4. Hasil Distribusi Label Sentimen.

Setelah proses *sampling*, diperoleh 5.000 data ulasan yang digunakan dalam penelitian, dengan distribusi sentimen yang didominasi oleh kelas positif. Kondisi ini menunjukkan adanya ketidakseimbangan kelas pada dataset berpotensi memengaruhi performa model pada kelas minor, khususnya sentimen netral dan *negative.*

Hasil *Preprocessing* Data

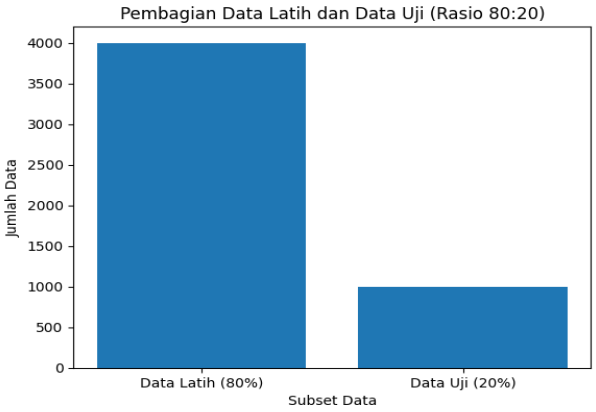
Tahap preprocessing menghasilkan teks ulasan yang bersih, terstandarisasi, dan distemming sehingga lebih ringkas serta fokus pada kata kunci sentimen, yang selanjutnya memberikan input optimal untuk ekstraksi fitur TF-IDF dan klasifikasi sentimen.

index	Teks Asli	Data Cleaning	Case Folding	Tokenizing	Normalization	Stopword Removal	Stemming	Clean Text Final
0	Terimakasih telah menyediakan transport yang sangat nyaman. Dan driver yang sangat baik. Dalijo. Sehingga kami sangat menikmati perjalannya. Next pasti saya pesan lagi	Terimakasih telah menyediakan transport yang sangat nyaman Dan driver yang sangat baik Dalijo Sehingga kami sangat menikmati perjalannya Next pasti saya pesan lagi	terimakasih telah menyediakan transport yang sangat nyaman dan driver yang sangat baik dalijo sehingga kami sangat menikmati perjalannya next pasti saya pesan lagi	terimakasih, telah, menyediakan, transport, yang, sangat, nyaman, dan, driver, yang, sangat, baik, dalijo, sehingga, kami, sangat, menikmati, perjalannya, next, pasti, saya, pesan, lagi	terimakasih, telah, menyediakan, transport, yang, sangat, nyaman, dan, driver, yang, sangat, baik, dalijo, sehingga, kami, sangat, menikmati, perjalannya, next, pasti, saya, pesan, lagi	terimakasih, menyediakan, transport, nyaman, driver, dalijo, menikmati, perjalannya, next, pesan	terimakasih sedia transport nyaman driver dalijo nikmat jalan next pesan	terimakasih sedia transport nyaman driver dalijo nikmat jalan next pesan
1	Lokasi bagus dan penjelasan tour guidenya sangat detail.. tapi ketat banget peraturannya. Jalan melencenh dikit ditegur, padahal saya lagi hamil dan bbrp waktu hrs duduk istirahat krn perjalanan tour lumayan jauh dan lama.	Lokasi bagus dan penjelasan tour guidenya sangat detail tapi ketat banget peraturannya Jalan melencenh dikit ditegur padahal saya lagi hamil dan bbrp waktu hrs duduk istirahat krn perjalanan tour lumayan jauh dan lama	lokasi bagus dan penjelasan tour guidenya sangat detail tapi ketat banget peraturannya jalan melencenh dikit ditegur padahal saya lagi hamil dan bbrp waktu hrs duduk istirahat krn perjalanan tour lumayan jauh dan lama	lokasi, bagus, dan, penjelasan, tour, guidenya, sangat, detail, tapi, ketat, banget, peraturannya, jalan, melencenh, dikit, ditegur, padahal, saya, lagi, hamil, dan, bbrp, waktu, hrs, duduk, istirahat, krn, perjalanan, tour, lumayan, jauh, dan, lama	lokasi, bagus, dan, penjelasan, tour, guidenya, sangat, detail, tapi, ketat, banget, peraturannya, jalan, melencenh, dikit, ditegur, padahal, saya, lagi, hamil, dan, bbrp, waktu, hrs, duduk, istirahat, karena, perjalanan, tour, lumayan, jauh, dan, lama	lokasi, bagus, penjelasan, tour, guidenya, detail, ketat, banget, peraturannya, jalan, melencenh, dikit, ditegur, hamil, bbrp, hrs, duduk, istirahat, perjalanan, tour, lumayan	lokasi bagus jelas tour guidenya detail ketat banget atur jalan melencenh dikit tegur hamil bbrp hrs duduk istirahat jalan tour lumayan	lokasi bagus jelas tour guidenya detail ketat banget atur jalan melencenh dikit tegur hamil bbrp hrs duduk istirahat jalan tour lumayan
2	Candi Prambanan, salah satu icon wisata yang berada di D.I. Yogyakarta, Candinya lebih kecil jika dibanding dengan Candi Borobudur tapi untuk Komplek Candi	Candi Prambanan salah satu icon wisata yang berada di D I Yogyakarta Candinya lebih kecil jika dibanding dengan Candi Borobudur tapi	candi prambanan salah satu icon wisata yang berada di d i yogyakarta candinya lebih kecil jika dibanding dengan candi borobudur tapi	candi, prambanan, salah, satu, icon, wisata, yang, berada, di, d, i, yogyakarta, candinya, lebih, kecil, jika, dibanding, dengan, candi, borobudur, tapi,	candi, prambanan, salah, satu, icon, wisata, yang, berada, di, d, i, yogyakarta, candinya, lebih, kecil, jika, dibanding, dengan, candi, borobudur, tapi,	candi, prambanan, salah, icon, wisata, d, i, yogyakarta, candinya, dibanding, candi, borobudur, komplek, candi, prambanan, luas, ramah, keluarga,	candi prambanan salah icon wisata d i yogyakarta candi banding candi borobudur komplek candi prambanan luas	candi prambanan salah icon wisata d i yogyakarta candi banding candi borobudur komplek candi prambanan luas ramah keluarga

Gambar 5. Hasil Akhir *Preprocessing* Data.

Pembagian Data

Setelah *preprocessing* data dilakukan, tahap selanjutnya adalah membagi data menjadi dua, yakni menjadi data latih dan data uji. Pada penelitian ini digunakan perbandingan sebesar 80 : 20. Berikut merupakan hasil pembagian data.



Gambar 6. *Bar Chart* Pembagian Data.

Hasil Ekstraksi Fitur dengan TF-IDF

Setelah data ulasan melalui tahap split data dilakukan *feature extraction* untuk mengubah data teks menjadi numerik agar dapat diproses oleh algoritma pembelajaran mesin. Pada penelitian ini digunakan metode Term *Frequency–Inverse Document Frequency* (TF-IDF) untuk merepresentasikan setiap dokumen dalam bentuk vektor berbobot.

Kode Program:

```
tfidf = TfidfVectorizer(
    max_features=800,
    min_df=5,
    max_df=0.9,
    sublinear_tf=True
)
X_train = tfidf.fit_transform(X_train_text)
X_test  = tfidf.transform(X_test_text)
print("Dimensi TF-IDF Train:", X_train.shape)
print("Dimensi TF-IDF Test :", X_test.shape)
```

Proses ekstraksi fitur menggunakan *TfidfVectorizer* dengan parameter sebagai berikut:

a. *max_features* = 800, untuk membatasi jumlah maksimum fitur kata
b. *min_df = 5*, untuk mengabaikan kata yang muncul pada kurang dari lima dokumen
c. *max_df = 0.9*, untuk menghilangkan kata yang muncul pada lebih dari 90% dokumen
d. *sublinear_tf* = True, untuk menerapkan skala logaritmik pada frekuensi kata

Parameter tersebut bertujuan untuk menghasilkan fitur yang lebih relevan serta mengurangi pengaruh kata-kata yang terlalu umum maupun terlalu jarang. Bedasarkan hasil ekstraksi fitur, diperoleh dimensi matriks TF-IDF sebagai berikut:

```
Dimensi TF-IDF Train: (4000, 800)
Dimensi TF-IDF Test : (1000, 800)
```

Gambar 7. Hasil Ekstraksi.

Data latih terdiri dari 4.000 dokumen dan data uji 1.000 dokumen dengan 800 fitur kata, di mana perhitungan rata-rata bobot TF-IDF digunakan untuk mengidentifikasi kata-kata yang paling berkontribusi dalam merepresentasikan dokumen pada data latih.

index	Kata	Bobot TF-IDF Rata-rata
0	jalan	0.03496165436142087
1	yg	0.03268806993162542
2	foto	0.029467865191256354
3	bagus	0.02934127307233167
4	candi	0.02853981960536822
5	wisata	0.028216798836892466
6	jogja	0.025609534551590967
7	pantai	0.025061615044444433
8	indah	0.02461589774213496
9	kunjung	0.02374891803551134

Nilai bobot diperoleh dari hasil perhitungan rata-rata TF-IDF pada data latih)

Gambar 8. Daftar Top 10 TF-IDF.

Kata-kata dengan bobot TF-IDF tertinggi menunjukkan tingkat kepentingan yang lebih besar dalam membedakan dokumen satu dengan lainnya. Hal ini disebabkan oleh tingginya frekuensi kemunculan kata pada dokumen tertentu dan rendahnya frekuensi kemunculan kata tersebut pada seluruh dokumen.Fitur-fitur kata ini kemudian digunakan sebagai input pada tahap pemodelan klasifikasi agar dapat meningkatkan performa model dalam mengenali pola pada data teks.

Analisis Sentimen

Pengujian performa model dilakukan menggunakan metode *Stratified K-Fold Cross Validation* dengan nilai k 10. Data akan melakukan pelatian setiap *fold* dari nilai k yang diberikan. Pendekatan ini digunakan untuk memastikan distribusi kelas sentimen tetap proporsional pada setiap *fold* dan hasil evaluasi lebih stabil.

Tabel 1. Hasil *K-Fold Cross Validation.*

Index	*Fold*	SVM	SVM + LDA
0	1	0.8525	0.88
1	2	0.8525	0.875
2	3	0.8525	0.88
3	4	0.8525	0.8775
4	5	0.8525	0.8925
5	6	0.8525	0.89
6	7	0.8525	0.875
7	8	0.8525	0.88
8	9	0.8525	0.8775
9	10	0.8525	0.9
Rata - rata		0.8525	0.8828

Hasil cross validation menunjukkan bahwa penerapan LDA sebagai metode reduksi dimensi mampu meningkatkan akurasi dan konsistensi kinerja model SVM dibandingkan tanpa LDA. Kombinasi SVM dan LDA menghasilkan akurasi rata-rata yang lebih tinggi serta pemisahan kelas yang lebih optimal, sementara evaluasi akhir dilakukan menggunakan pembagian data latih dan uji 80:20 dengan metrik akurasi, presisi, recall, dan F1-score.

Tabel 2. Evaluasi Model SVM.

Index	*Precission*	*Recall*	*F1-score*	*Accuracy*
Negatif	0.0	0.0	0.0	0.853
Netral	0.0	0.0	0.0	0.853
Positif	0.853	1.0	0.9207	0.853

Model SVM hanya mampu mengklasifikasikan kelas *positive* dengan baik, ditunjukkan oleh nilai *recall* sebesar 1.0 dan *f1-score* 0.9207. Pada kelas negatif dan netral tidak terdeteksi sama sekali, karena nilai *precision, recall,* dan *f1-score* pada kedua kelas tersebut bernilai 0. Meskipun nilai akurasi mencapai 85,3%, hasil ini bersifat menyesatkan karena diperoleh dari prediksi satu kelas saja. Artinya performa model belum optimal untuk klasifikasi multikelas. Selanjutnya pada model SVM dengan LDA menghasilkan evaluasi yang berbeda

Tabel 3. Evaluasi Model SVM + LDA.

Index	*Precission*	*Recall*	*F1-score*	*Accuracy*
Negatif	0.1667	0.0833	0.1111	0.821
Netral	0.2571	0.1333	0.1756	0.821
Positif	0.868	0.9402	0.9026	0.821

Hasil akurasi model SVM dengan LDA sebesar 82,1%, secara akurasi lebih rendah dari model sebelumnya. Nilai akurasi ini menunjukkan kinerja model yang cukup baik secara keseluruhan. Namun model masih cenderung bias ke kelas positif sehingga perlu perbaikan lebih lanjut untuk menyeimbangkan performa antar kelas. Model SVM + LDA menunjukkan

peningkatan kemampuan dalam mengenali kelas negatif dan netral, meskipun nilainya masih relatif rendah. Berikut hasil perbandingan metrik evaluasi tiap model pada Tabel 4.

Tabel 4. Perbandingan Hasil Evaluasi.

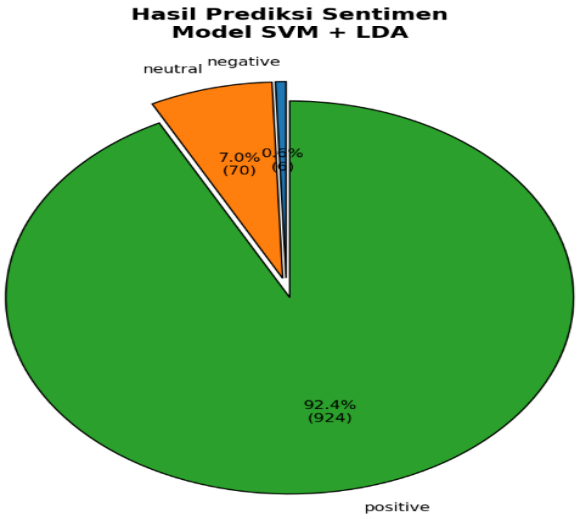
Index	*Model*	*Precission*	*Recall*	*F1-score*	*Accuracy*
0	SVM	0.7276	0.853	0.7853	0.853
1	SVM + LDA	0.7771	0.821	0.795	0.821

Analisis metrik menunjukkan bahwa model SVM dengan LDA memiliki nilai precision lebih tinggi dibandingkan SVM tanpa LDA, yang menandakan peningkatan ketepatan klasifikasi dan penurunan kesalahan positif palsu.

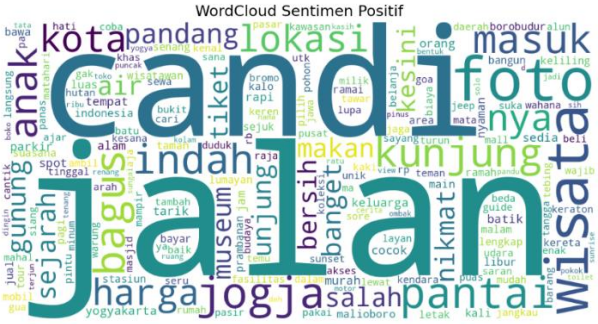
Model SVM tanpa LDA menunjukkan nilai recall dan accuracy yang lebih tinggi, sedangkan kombinasi SVM + LDA menghasilkan nilai F1-score yang sedikit lebih unggul, yang mengindikasikan bahwa penggunaan LDA mampu memberikan keseimbangan yang lebih baik antara precision dan recall meskipun akurasi keseluruhan sedikit menurun.

Model SVM tanpa LDA unggul pada recall dan accuracy, sedangkan SVM dengan LDA meningkatkan precision dan F1-score, sehingga model SVM dengan LDA dipilih karena menghasilkan klasifikasi sentimen yang lebih seimbang dan presisi meskipun dengan sedikit penurunan kinerja lainnya.

Selanjutnya model terbaik berupa kombinasi SVM dan LDA digunakan untuk memprediksi sentimen ulasan wisata di Kota Yogyakarta dengan mengklasifikasikan ulasan ke dalam sentimen positif dan negatif, di mana LDA berperan dalam reduksi dimensi dan SVM sebagai pengklasifikasi utama untuk menghasilkan pemetaan persepsi wisatawan yang lebih akurat.

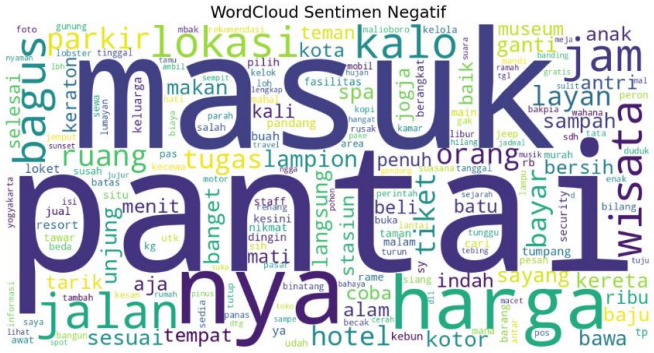


Gambar 9. Hasil Prediksi Sentimen.



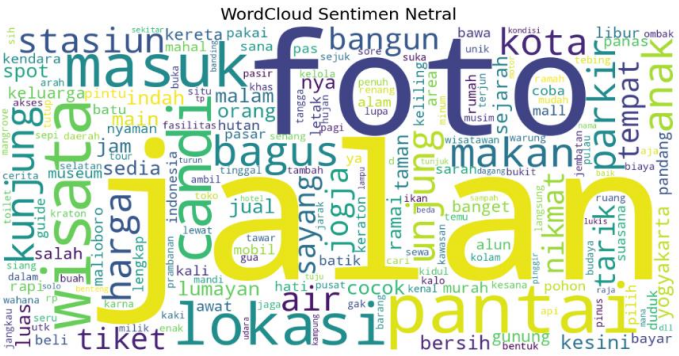
Gambar 10. *Wordcloud* Sentimen Positif.

Pada Gambar 10, frekuensi kata pada sentimen positif yang paling sering digunakan, “bagus”, “lokasi”, “indah”, “nyaman”, “murah”, harga, “pemandangan” dan “bersih”. Dominasi kata-kata tersebut menunjukkan bahwa ulasan positif wisata di Yogyakarta banyak dipengaruhi oleh aspek keindahan, kenyamanan, kebersihan, kemudahan lokasi, serta keterjangkauan harga destinasi wisata.



Gambar 11. *Wordcloud* Sentimen Negatif.

Pada sentimen negatif, beberapa kata dengan frekuensi tinggi yaitu, “masuk”, “pantai”, “mahal”, “jalan”, “parkir”, “ramai”, “kotor”, “panas”, dan “antri”. Dominasi kata-kata tersebut menunjukkan bahwa sentimen negatif pada ulasan wisata di Yogyakarta umumnya dipicu oleh permasalahan akses dan fasilitas, kepadatan pengunjung, kondisi lingkungan, serta tingkat kenyamanan yang dirasakan wisatawan.



Gambar 12. *Wordcloud* Sentimen Netral.

Pada *WordCloud* sentimen netral, kata yang sering digunakan meliputi "jalan", "lokasi", "harga", "masuk", "pantai", "makan", dan "stasiun". Kemunculan kata-kata ini menunjukkan bahwa ulasan netral cenderung bersifat informatif dan deskriptif berfokus pada penyampaian informasi umum mengenai lokasi, akses, dan fasilitas tanpa memberikan penilaian emosional yang kuat.

5. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian ini menganalisis 5.000 ulasan wisata di Kota Yogyakarta menggunakan TF-IDF dan model SVM, di mana penerapan LDA mampu meningkatkan keseimbangan performa klasifikasi melalui nilai precision dan F1-score yang lebih baik. Hasil klasifikasi menunjukkan dominasi sentimen positif, sehingga membuktikan efektivitas kombinasi preprocessing teks, TF-IDF, serta SVM dengan LDA dalam memprediksi persepsi wisatawan.Hasil klasifikasi SVM + LDA menunjukkan dominasi sentimen positif pada ulasan wisata di Yogyakarta, didukung analisis wordcloud, sementara sentimen negatif dan netral berkaitan dengan aspek fasilitas dan informasi, sehingga membuktikan efektivitas kombinasi preprocessing, TF-IDF, serta SVM dengan LDA dalam memprediksi persepsi wisatawan.

Saran

Berdasarkan temuan penelitian ini, disarankan kepada pengelola objek wisata di Kota Yogyakarta untuk mempertahankan dan meningkatkan aspek-aspek yang mendapatkan respons positif dari wisatawan, seperti keindahan lingkungan, kenyamanan, kebersihan, kemudahan akses lokasi, serta keterjangkauan harga. Di sisi lain, perhatian khusus perlu diberikan pada faktor-faktor yang memicu sentimen negatif, terutama terkait perbaikan akses menuju lokasi wisata, peningkatan kualitas dan ketersediaan fasilitas, pengelolaan kepadatan pengunjung, serta pemeliharaan kondisi lingkungan agar tetap nyaman dan berkelanjutan.

Selain itu, bagi penelitian selanjutnya disarankan untuk memperluas jumlah dan variasi sumber data ulasan, tidak hanya terbatas pada TripAdvisor tetapi juga platform lain seperti Google Reviews atau media sosial, sehingga hasil analisis dapat lebih komprehensif. Penggunaan metode klasifikasi lain atau pendekatan *deep learning*, serta penambahan analisis temporal sentimen, juga dapat dipertimbangkan untuk memperoleh pemahaman yang lebih mendalam mengenai dinamika persepsi wisatawan dari waktu ke waktu.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam pelaksanaan serta penyusunan penelitian ini. Ucapan terima kasih disampaikan kepada PT. Global Data Inspirasi yang telah menyediakan fasilitas dan sumber data yang digunakan dalam penelitian ini.

DAFTAR REFERENSI

Ain, Q., Utami, E., & Nasiri, A. (2024). Analisis sentimen: Prediksi. *Jurnal Ilmiah*, 9(3), 1586–1595.

Almaspoor, M. H., Safaei, A., Salajegheh, A., & Minaei-Bidgoli, B. (2021). *Support vector machines in big data classification: A systematic literature review*. ResearchSquare, 1–34.

Aryanto, P. M., & Mardhiyyah, R. (2024). Analisis sentimen terhadap review Google Maps Jogja City Mall menggunakan algoritma support vector machine. *Journal of Computer System and Informatics (JoSYC)*, 6(1), 25–35. https://doi.org/10.47065/josyc.v6i1.6045

Chetna, K. (2025). Sentiment analysis using SVM classifier in data mining: A machine learning approach. *International Journal of Computer Science & Information Technology*, 1(1), 29–33.

Diandra, D. (2022). *Analisis sentimen ulasan MyXL dengan support vector machine* [Skripsi/tesis tidak dipublikasikan].

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An overview on the advancements of support vector machine models in healthcare applications: A review. *Information*, 15(4). https://doi.org/10.3390/info15040235

Hanafi, M., Ridwan, M., & Nooriansyah, S. (2025). Analisis sentimen pengunjung terhadap objek wisata Kabupaten Gresik menggunakan support vector machine (SVM) dan linear discriminant analysis (LDA). *Jurnal Ilmu Komputer dan Desain Komunikasi Visual*, 10(1), 91–107.

Ipmawati, J., Saifulloh, S., & Kusnawi, K. (2024). Analisis sentimen tempat wisata berdasarkan ulasan pada Google Maps menggunakan algoritma support vector machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(1), 247–256.

Larasati, L. L. (2024). *Analisis sentimen tentang wisata di Yogyakarta pada platform Instagram menggunakan naïve Bayes classifier* [Skripsi/tesis tidak dipublikasikan].

Lestari, V. B., & Hutagalung, C. A. (2025). Evaluation of TF-IDF extraction techniques in sentiment analysis of Indonesian-language marketplaces using SVM, logistic regression, and naïve Bayes. *J-KOMA: Journal of Computer Science and Applications*. https://doi.org/10.21009/j-

Syahlan, M. S., Irmayanti, D., & Alam, S. (2023). Analisis sentimen terhadap tempat wisata dari komentar pengunjung dengan menggunakan metode support vector machine. *Simtek: Jurnal Sistem Informasi dan Teknik Komputer*, 8(2), 315–319. https://doi.org/10.51876/simtek.v8i2.281

Tri Putra, K., Hariyadi, M. A., & Crysdian, C. (2021). Perbandingan feature extraction TF-IDF dan bag of words untuk analisis sentimen berbasis SVM. *Jurnal Cahaya Mandalika*, 1449–1463.

Wahyudi, T., Rudiman, & Verdikha, N. A. (2024). Klasifikasi sentimen X (Twitter) perihal pemindahan ibu kota Indonesia menggunakan ekstraksi fitur TF-IDF dan metode support vector machine (SVM). *JTI: Jurnal Teknologi Informasi*, 18(2), 185–199.

Winarnie, W., Kusrini, K., & Hartanto, A. D. (2023). Pengurangan dimensi dengan metode linear discriminant analysis (LDA). *Infotek: Jurnal Informatika dan Teknologi*, 6(2), 228–237. https://doi.org/10.29408/jit.v6i2.10069

Zhahrina, A., Sofiah, U., Wahyu, D., Andayani, A., Khairurrabbani, C., & Saputri, F. (2025). Penerapan metode support vector machine (SVM) dalam analisis sentimen ulasan aplikasi Tokopedia di Google Play Store. *Prosiding/Jurnal Ilmiah*, 84–91.

Zulfikar, A. F. (2017). Pengembangan algoritma stemming bahasa Indonesia dengan pendekatan dictionary-based stemming untuk menentukan kata dasar dari kata yang berimbuhan. *Jurnal Informatika Universitas Pamulang*, 2(3), 143. https://doi.org/10.32493/informatika.v2i3.1443