Jurnal Ilmu Komputer dan Desain Komunikasi Visual Volume 10, Nomor 1, Juli 2025
ISSN (Cetak): 2541-4550 ISSN (Online): 2541-4585

Analisis Sentimen Pengunjung terhadap Objek Wisata Kabupaten Gresik Menggunakan *Support Vector Machine* (SVM) dan *Linear Discriminant Analysis* (LDA)

Muhammad Hanafi[1], Mujib Ridwan[2], Subhan Nooriansyah[3]

[1,2,3]*Program Studi Sistem Informasi, Fakultas Sains dan Teknologi,*
UIN Sunan Ampe1 Surabaya
Jalan Dr. Ir. H. Soekarno No.682, Surabaya
[1]hanafim3000@gmail.com, [2]mujibrw@uinsby.ac.id, [3]subhan.nooriansyah@uinsby.ac.id

Abstrak

Sektor pariwisata di pulau Jawa mendominasi arus perjalanan domestik di Indonesia. Jawa Timur menyumbang angka tertinggi dengan 198,91 juta perjalanan. Namun, kondisi ini masih belum merata ke seluruh daerah. Berdasarkan data kunjungan wisatawan online (DAKUWISON) tercatat bahwa terjadi penurunan pengunjung wisata di Kabupaten Gresik pada tahun 2023. Hal ini tidak sesuai dengan kebijakan PPKM yang ditiadakan pada tahun sebelumnya. Penelitian ini bertujuan menganalisis sentimen ulasan menggunakan metode klasifikasi SVM-LDA untuk mengetahui persepsi mereka sebagai tambahan opini berbasis data bagi pengelola wisata. Support Vector Machine (SVM) sebagai metode Supervised Learning diterapkan dalam penelitian, selain itu peningkatan klasifikasi dengan menambahkan metode reduksi dimensi Linear Discriminant Analysis (LDA). Pengambilan data dari Google Maps dengan teknik web scrapping diperoleh 3460 ulasan. Hasil dari penelitian dari perbandingan evaluasi masing-masing model menunjukkan bahwa model SVM dengan LDA dapat mengungguli dari model SVM yang tidak menerapkan LDA. Nilai fI-score dari model SVM dengan LDA lebih tinggi di angka 66% dibandingkan dengan model SVM yang tidak menerapkan LDA dengan nilai fI-score 53%. Berdasarkan hasil klasifikasi sentimen pada data 2023 menunjukkan bahwa sentimen pengunjung cenderung positif dari 511 ulasan diperoleh 456 sentimen positif, 33 sentimen negatif, dan 22 sentimen netral.

Kata.kunci— *Ana1isis Sentimen, Ulasan, Wisata, Support Vector Machine, Linear-Discriminant Ana1ysis*

Abstract

The tourism sector on the island of Java dominates the flow of domestic travel in Indonesia. East Java contributed the highest number with 198.91 million trips. However, this condition is still not evenly distributed across all regions. Based on the Online Tourist Visit Data (DAKUWISON), it was noted that there was a decrease in tourist visitors in Gresik Regency in 2023. This is not following the PPKM policy that was eliminated in the previous year. This research purposes to analyze the sentiment of reviews using the SVM-LDA classification method to determine their perceptions as an additional data-based opinion for tourism managers. Support Vector Machine (SVM) as a Supervised Learning method is applied in research, besides improving classification by adding the Linear Discriminant Analysis (LDA) dimension reduction method. Data collection from Google Maps with a web scraping technique obtained 3460 reviews. The results of research from the evaluation comparison of each model show that the SVM model with LDA is better than the SVM model without LDA. The f1-score value of the SVM model with LDA is 66% higher than the SVM model without LDA, with an f1-score value of 53%. Based on the results of sentiment classification on 2023 data, it shows that visitor

sentiment tends to be positive, with 511 reviews, 456 positive sentiments, 33 negative sentiments, and 22 neutral sentiments obtained.

Keywords— *Sentiment Ana1ysis, Reviews, Tourism, Support Vector Machine, Linear Discriminant Ana1ysis*

1. PENDAHULUAN

Indonesia merupakan salah satu wilayah di Asia yang memiliki keragaman sumber daya alam dan ekosistem yang melimpah. Dalam beberapa sektor industri sangat terbantu dengan keragamannya, sehingga peningkatan perekonomian dapat dirasakan terutama pada sektor pariwisata terutama pada sektor pariwisata (Utami & Erfina, 2022). Berdasarkan data Badan Pusat Statistik (BPS) pada tahun 2022, Pulau Jawa mendominasi arus perjalanan pariwisata domestik di Indonesia. Tercatat angka tertinggi dengan 198,91 juta perjalanan pada provinsi Jawa Timur mengungguli provinsi Jawa Barat dan Jawa Tengah (Santika, 2023).

Tidak hanya BPS, terdapat data yang lain yaitu Produk Domestik RegionaI Bruto (PDRB), posisi kedua dihuni oleh Provinsi Jawa Timur dalam hal PDRB terkaya setelah DKI Jakarta, dengan total PDRB sekitar Rp 2.454.499 miliar, sementara jumlah penduduknya mencapai sekitar 40.878.800 orang. Salah satu wilayah di provinsi tersebut, yakni Kabupaten Gresik memiliki pendapatan per kapita tahunan tertinggi sebesar 109.313.000, mengungguli 29 kabupaten lain di Jawa Timur. (Saputra, 2022).

Selain itu, bukti dari sumber lain yaitu Data Kunjungan Wisata *Online* (DAKUWISON) (Dakuwison, 2018) di tahun 2022, kondisi arus perjalanan pengunjung wisata di Kabupaten Gresik dalam keadaan stabil dengan rata-rata 300-350 pengunjung setiap bulan dari wisatawan nusantara atau mancanegara. Namun kondisi tersebut tidak bertahan lama. Pada tahun 2023, pengunjung wisata mengalami penurunan di setiap bulannya. Pada bulan April tercatat hanya 100 ribu pengunjung. Peraturan pemerintah mengenai Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) telah ditiadakan sejak akhir tahun 2022. Namun hal tersebut tidak menjadi perubahan terhadap pengunjung wisata Kabupaten Gresik yang menurun dibanding tahun sebelumnya.

Faktor lainnya terdapat pada ulasan pengunjung lain yang sebelumnya telah berkunjung ke wisata pilihannya, dan membuat penilaian terhadap wisata tersebut. Informasi tersebut dapat dijadikan oleh pengunjung baru terkait bagaimana kondisi wisata yang akan dikunjunginya (Herlawati dkk., 2021). Informasi rating dan ulasan umumnya dapat diperoleh dari *platform* Google Maps. Aplikasi ini menjadi aspek penting dalam era big data untuk memperoleh informasi tersebut (Haq, 2020). Namun terkadang terdapat beberapa ulasan atau komentar yang tidak sesuai dengan *star rating* yang diberikan. Pengunjung atau penulis dapat secara bebas memberikan penilaian, yang secara otomatis muncul notifikasi dari Google Maps (Hesay dkk., 2021).

Teknik yang dapat diimplementasikan dalam menganalisis data dengan skala besar salah satunya analisis sentimen. Ulasan pengunjung lewat Google Maps digunakan sebagai data yang dianalisis dengan tujuan memberikan jawaban pola sudut pandang para pengunjung terhadap objek wisata di Kabupaten Gresik. Kemudian ulasan diproses untuk menghasilkan klasifikasi sentimen diperuntukkan kepada para pengelola wisata sebagai solusi tambahan dalam menentukan keputusan yang sesuai untuk proses peningkatan fasilitas atau infrastruktur di wisata masing-masing.

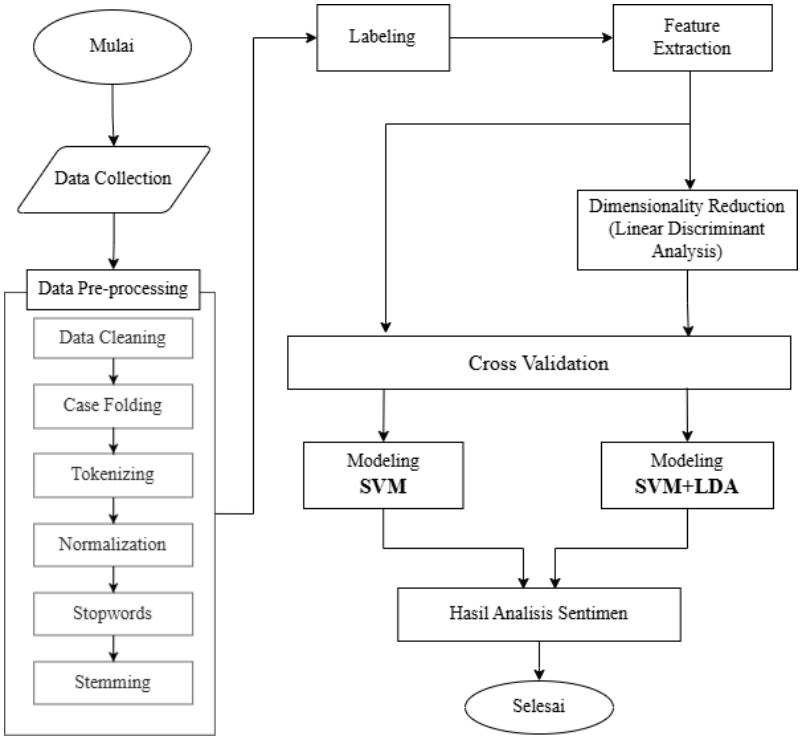
Pratama, dkk (2018) melakukan penelitian dengan tujuan untuk menganalisa aspek dari opini pengunjung wisata pantai Malang menggunakan analisis sentimen. Sebanyak 43 objek pantai di Malang Selatan digunakan dari sumber *TripAdvisor*. Evaluasi SVM menghasilkan *Accuracy* sebesar 87%, dan berhasil melakukan klasifikasi pada tiap aspek. Hasil *Usability testing* dari *Dashboard* memperoleh nilai 70 dan termasuk dalam kategori *Acceptable* dengan *rating Good*. Penelitian Prasetyo dan Hidayatullah (Prasetyo & Hidayatullah, 2020), mengidentifikasi dual sentimen dari ulasan objek wisata di Yogyakarta, dengan model *Logistic Regression, Naïve Bayes Classifier* dan *Support Vector Machine*, menghasilkan performa terbaik pada model SVM dalam memprediksi kalimat *dual sentiment* dengan akurasi 83%.

Berdasarkan beberapa penelitian yang telah dilakukan, bahwa metode SVM memberikan hasil evaluasi yang baik dengan kasus klasifikasi sentimen dalam topik yang berbeda. Namun pada penelitian Yue, penggunaan SVM memiliki kelemahan seperti waktu pelatihan atau *training* yang lama sehingga membuat biaya komputasi menjadi mahal. Penulis memberikan tambahan dengan menambahkan metode *Dimensionality Reduction*, karena pada kasus penelitian sebelumnya (Zebari dkk., 2020), dapat mereduksi kompleksitas waktu dan pemakaian memori. *Linear Discriminant Analysis* digunakan sebagai metode *Dimensionality Reduction*.

Pada penelitian ini terdapat beberapa tahapan, diawali dengan pengumpulan data, dan dilanjut data pre-processing. Pelabelan data menggunakan *TextBlob* dengan validasi dari pengelola wisata. Dalam proses vektorisasi kata menggunakan teknik word embedding dengan *FastText*. Pada tahap klasifikasi menggunakan algoritma *Support Vector Machine* (SVM) karena memiliki hasil akurasi yang baik. Namun pada model SVM memiliki kekurangan dari sisi penggunaan memori yang besar dan biaya. Sebagai bahan pertimbangan, penulis menambahkan metode dimensionality reduction dengan *Linear Discriminant Analysis* sebagai solusi untuk masalah kelemahan dari SVM. *K-Fold Cross Validation* akan digunakan untuk mengetahui hasil evaluasi training model. Penelitian ini bertujuan menganalisis sentimen yang berdasar dari ulasan pengunjung di wisata di Kabupaten Gresik 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 wisata.

2. METODE PENELITIAN

Metode penelitian ini menerapkan kuantitatif deskriptif yang mana data dikumpulkan dan dianalisis sentimen ulasan atau *review* pengunjung. Data yang digunakan berfokus pada ulasan objek wisata Kabupaten Gresik di *Google Maps* dengan *rating* 3 ke atas dan lebih dari 100 ulasan. Diawali pengumpulan data, data *pre-processing*, labeling (menentukan label*)*, *feature extraction*, *dimensionality reduction,* dan proses evaluasi model. Beberapa langkah-langkah dalam penelitian sesuai alur pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Data Collection

Teknik *web scrapping* dilakukan pada *Google Review* dengan bantuan pustaka *Selenium* yang tersedia dalam *Python*. Ulasan yang diambil hanya pada objek wisata Kabupaten Gresik sesuai kriteria dan berkisar antara Mei 2020 sampai Mei 2023. Pemilihan ini dilakukan karena pada tahun 2020 kasus pandemi *covid-19* baru terjadi dan antara pada tahun 2022 – 2023, *covid-19* sudah mereda. Berdasarkan DAKUWISON (Dakuwison, 2018) per tahun 2022-2023, kondisi arus perjalanan pengunjung wisata di Kabupaten Gresik mengalami penurunan. Hal tidak sejalan dengan kebijakan PPKM yang telah ditiadakan dari tahun 2022. Total seluruh data terkumpul sebanyak 3460 data. Rincian data setiap tahun dalam Tabel 1. Ulasan pada tahun 2022 lebih banyak dibanding tahun lainnya. Hal tersebut terjadi karena kasus Covid-19 mulai mereda dan masyarakat mulai melaksanakan aktivitas seperti biasa.

Tabel 1. Data Terkumpul

	Periode			
	2020	**2021**	**2022**	**2023**
Data	338	736	2046	340
Total	**3460**			

2.2. Data Pre-processing

Data *Pre-processing* adalah serangkaian proses dalam mempersiapkan data dengan membersihkan data, mengubah data, dan mereduksi data agar data lebih relevan saat pengolahan data tahap selanjutnya (Larasati dkk., 2022). Pada tahap ini dilakukan proses eksplorasi data untuk mengetahui duplikasi data dan kolom kosong. Hasil setelah eksplorasi data seperti pada Tabel 2. Total data sedikit berkurang dari proses ini.

Tabel 2. Setelah Eksplorasi

	Periode			
	2020	**202l**	**2022**	**2023**
Data	262	648	1872	330
Total	**3112**			

Tahapan data *pre-processing* menyesuaikan dengan relevansi kasus yang dianalisis, berikut beberapa tahapannya:

2.2.1 Cleaning

Proses awal dilakukan pembersihan terhadap data dengan menghapus beberapa komponen pendukung kalimat seperti simbol, tanda baca, angka, *double space*, dan emotikon. Atribut yang dihilangkan atau dihapus memang tidak berkaitan dengan tahap proses pengolahan data.

2.2.2 Case Folding

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

2.2.3 Tokenizing

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

2.2.4 Normalizaation

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

2.2.5 Remove Stopwordsas

Proses berlanjut ke *remove stopwords*. Kata-kata yang tidak bermakna atau tidak berhubungan dengan kasus akan dihilangkan. Proses ini dibantu dengan pustaka NLTK dengan kamus *stopwords* yang sudah tersedia. Penambahan pada kamus *stopwords* dilakukan seperti kata *"masszeh", "cak", "cok", "tok", "mashok"* dan kata lain yang tidak memiliki sentimen.

2.2.6 Stemming

Proses terakhir, mengubah kata dengan imbuhan ke kata dasar berdasarkan kamus. Pada tahap *stemming*, pustaka kamus bahasa Indonesia dari sastrawi pada *Python*. Setelah rangkaian proses dilakukan, akan membuat data lebih mudah dalam proses pengolahan berikutnya.

2.3. Labeling

Data yang telah melewati *pre-processing*, selanjutnya diberikan label guna menentukan jenis data berlabel positif, negatif, atau netral. Setiap data teks yang sebelumnya terbagi menjadi token akan digabungkan kembali untuk proses pelabelan. Proses dibantu pustaka *TextBlob* dengan menghitung nilai polaritas setiap kata. Penambahan kamus leksikon bahasa Indonesia (Anasta, 2023) dilakukan agar penilaian polaritas lebih sesuai. Rumus penentuan nilai polaritas sebagai berikut.

$$Polaritas = \frac{w_1 + w_2 + \ldots . w_n}{s_n} \qquad (1)$$

Nilai rata-rata polaritas diperoleh dari nilai polaritas setiap kata (w_n) dibagi dengan jumIah kata (s_n) pada teks. Hasil nilai tersebut akan digunakan sebagai tanda penentuan jenis label dari setiap teks pada data dengan kondisi sesuai Tabel 3.

Tabel 3. Acuan Polaritas

Polaritas	Label
>0	Positif
0	Netral
<0	Negatif

Hasil pelabelan data akan divalidasi dengan label berdasarkan dari penilaian pengelola wisata. Validasi dilakukan untuk mengetahui tingkat keakuratan antara mesin dan manusia dalam memberikan keputusan. Penelitian Lai dan Tan (Lai & Tan, 2019), menyatakan tingkat kecocokan mesin dan manusia antara 70-80%. Data yang digunakan dalam validasi berupa data sampel dengan teknik *judgement sampling*. Teknik ini tergolong dalam teknik *non-probability sampling* dan dapat diterapkan pada kasus analisis sentimen (Lappeman dkk., 2020). Proses validasi melibatkan 3 pengelola wisata yang berbeda. Pengelola akan memberikan jawaban berdasarkan sudut pandang mereka terkait sentimen dalam ulasan. Hasil validasi dan pelabelan *TextBlob* akan dihitung tingkat akurasi kesesuaian label.

2.4. Feature Extraction

Proses *feature extraction* sangat penting dilakukan dalam *machine learning*. *Feature extraction* merupakan proses mengekstrak fitur-fitur penting dalam data, berupa nilai (angka) (Cahyanti dkk., 2020). *Word embedding* sebagai salah satu teknik *feature extraction* berperan penting dalam menghasilkan nilai vektor dengan cara mengambil informasi semantik dari kata-kata dalam mengukur kesamaan kata (ÇeliK & Koç, 2021).

2.4.1 FastText

FastText merupakan pustaka dalam mempelajari teknik *word embedding* yang dikembangkan oleh *Facebook*. *FastText* umumnya digunakan dalam kasus klasifikasi kalimat dan representasi kata (Agustiningsih dkk., 2022). Rumus *FastText* dapat dipresentasikan pada rumus (2).

$$\sum_{t=1}^{T}\left[\sum_{c\in C_t}\ell\left(s(w_t.w_c)\right)+\sum_{n\in\mathcal{N}_{t.c}}\ell\left(-s(w_t.n)\right)\right] \qquad (2)$$

Lambang dalam rumus tersebut memilik arti simbol *s* berarti nilai skor, bobot dengan simbol *w*, simbol *l* berasal dari *log(1 +-e − x),* dan jumIah kata yang terkandung pada korpus atau *n*. Setiap kata akan memiliki vektor dan tersimpan dalam model *word embedding.*

Model *word embedding* yang telah dilatih memiliki panjang dimensi vektor 300 dan menghasilkan 3986 *vocabulary*. Setiap *vocabulary* dari kata-kata yang tersimpan dalam model memiliki angka vektor masing-masing. Hasil vektor ditampilkan pada Tabel 4.

Tabel 4. Dimensi Vektor

Vocabulary	**Vektor**
bagus	[-0.01574064 -0.131223 -0.05647508 -0.2058455...]

Vocabulary	**Vektor**
sesuai	[-0.03.91321 -0.10233811 -0.0719699 -0.22730872..]
mahal	[-0.0162349 -0.10825755 -0.07638002 -0.2466729..]

Dimensi dari setiap *vocabulary* pada model adalah 300. Model ini akan digunakan dalam membuat vektorisasi dari data *pre-processing* yang digunakan pada data fitur pembuatan model *machine learning*.

2.4.2 Linear Discriminant Analysis

Data yang telah berbentuk vektor memiliki jumlah dimensi yang cukup besar. Pada proses *dimensionality reduction*, setiap dimensi yang dimiliki setiap fitur akan direduksi sesuai nilai *n_component*. Pendekatan reduksi dimensi yang umum digunakan dalam data *mining* dan *machine learning* sebagai tahapan pra-pemrosesan disebut *Linear Discriminant Analysis* (LDA) (Reddy dkk., 2020). Proses ini tidak menghilangkan informasi yang terdapat pada kelas. Penelitian ini memiliki 3 kelas, nilai *n_component* yang digunakan dua. PemiIihan nilai *n_component* berdasarkan dari *n-kelas* dikurangi 1 (*n-kelas* – 1).

2.5. Klasifikasi SVM

Pembentukan model menggunakan algoritma SVM menggunakan parameter optimal yang diperoleh lewat proses *Cross Validation.* Pembagian data menjadi data latih atau fitur dengan lambang X_i dan nilai target atau label y_i (Pratama dkk., 2018). Konsep klasifikasi SVM berfokus pada pencarian *hyperplane* atau garis pemisah yang dapat memaksimalkan margin antar kelas. Kemampuan SVM dalam menemukan *hyperplane* yang optimal membuatnya memiliki tingkat generalisasi yang baik, sehingga berkontribusi pada peningkatan akurasi klasifikasi. *Classifier* akan terbentuk dengan persamaan (3) berikut:

$$f(x_i) = \{\geq 0, y_i = +1, < 0, y_i = -1\} \quad (3)$$

Hyperplane diperoleh dengan rumus (4) berikut:

$$W.x + b = 0 \quad (4)$$

Dengan *W* sebagai nilai bobot *support vector*, *b* sebagai nilai bias, *X* yang berarti data latih.

Keterangan:
W: nilai dalam vektor
b : nilai bias
X : data latih

2.5.1 Cross Validation

Proses berlanjut ke pemberlakuan *cross validation*. Salah satu metodenya dengan *k-fold cross validation.* Penerapan proses ini untuk menghindari *overfitting* saat pelatihan model. Proses validasi akan berlangsung selama dengan pemberian nilai *k* yang diberikan. Dalam tahap ini dilakukan *hyperparameter tuning* untuk menentukan parameter yang optimal. Pustaka *GridSearchCV* digunakan saat proses berlangsung.

Algoritma SVM bekerja dengan membuat garis linear atau *hyperplane* antara dua kelas berbeda. Pada data multi kelas atau non-linear perlu ditambahkan fungsi *kernel*

(Suryawan dkk., 2023). Berikut parameter-parameter yang digunakan dalam *hyperparameter tuning.*

Tabel 5. Paremeter *Tuning*

	Parameter		
	Kerne1	***C***	***Gamma***
Va1ue	*RBF, Sigmoid*	1, 10, 100	10, 1, 0,1

2.5.2 Modeling

a. Training

Tahap pertama data yang berupa vektor akan dibagi dengan rasio 80:20. Dari seluruh data yang telah melalui proses vektorisasi, sebanyak 80% data dialokasikan untuk data latih dan 20% untuk data uji. Data berupa vektorisasi angka yang diperoleh dari model *word embedding* sebelumnya. Dalam tahap ini, terdapat dua model yang akan buat. Model SVM (tidak menggunakan LDA), dan model SVM dengan reduksi dimensi LDA.

b. Evaluasi

Paska model dibuat akan dievaluasi dengan beberapa metrik dari *Confusion Matrix.* Tabel yang terbentuk dari *Confusion Matrix* mewakili setiap komponen antara label faktual dan prediksi. Beberapa komponen tersebut diantaranya, *True Positiveo*(TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Perhitungan evaluasi metrik akan berpengaruh terhadap angka yang dihasilkan setiap komponen pada *Confusion Matrix*

Precision memberikan gambaran kecocokan dari data terhadap prediksi hasil dari model. Nilai diperoleh dari persamaan (5) berikut.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall adalah rasio perbandingan antara data klasifikasi terhadap jumlah data yang relevan, dihitung dari persamaan (6) berikut

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F-measure atau *F1-score* adalah rata-rata harmonik antara nilai *Precision* dan *RecaII*. Persamaan (7) berikut dalam menentukan nilai *F1-Score*

$$F1\text{-}Score = 2\frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Accuracy adalah pengukuran nilai keakuratan model dalam melakukan klasifikasi. Semakin besar nilai *Accuracy*, hasil klasifikasi mendekati akurat dengan mempertimbankan metrik lainnya. Persamaan (8) untuk menghitung nilai *Accuracy*.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

Model yang selesai dibuat akan digunakan untuk memprediksi data baru. Data tersebut diambil pada ulasan hanya pada tahun 2023. Pemilihan model berdasarkan perbandingan hasil metrik evaluasi yang lebih baik. Model yang memilik hasil yang lebih optimal tanpa adanya kendala digunakan sebagai model klasifikasi terkait sentimen terhadap objek wisata di Kabupaten Gresik

3. HASIL DAN PEMBAHASAN

3.1. Data Penelitian

Setelah data dieksplorasi, langkah berikutnya data *pre-processing.* Terdapat 6 tahapan diantaranya *cleaning, cas folding, tokenizing, normalization, remove stopwords,* dan *stemming.* Hasil processing terdapat pada beberapa Tabel 6 dan 7.

Tabel 6. Hasil Data *Pre-Processing* I

Ulasan	*Cleaning*	*Case folding*
Good, Semakin banyak permainan Siiip..👍	Good Semakin banyak permainan Sip	good semakin banyak permainan sip
Wisatanya bersih, rekomendasi berkunjung pada saat pagi dan sore👍👍👍👍👍 …	Wisatanya bersih rekomendasi berkunjung pada saat pagi dan sore	wisatanya bersih rekomendasi berkunjung pada saat pagi dan sore
Menjadi alternatif eduwisata yg ada Gresik, semoga tambah baik	Menjadi alternatif eduwisata yg ada Gresik semoga tambah baik	menjadi alternatif eduwisata yg ada gresik semoga tambah baik
Biasa sj...panas pengelolaan ny krg	Biasa sj panas pengelolaan ny krg	biasa sj panas pengelolaan ny krg

Hasil berikutnya proses *tokenizing, normalization, remove stopwords, stemming.*

Tabel 7. Hasil Data *Pre-Processing* II

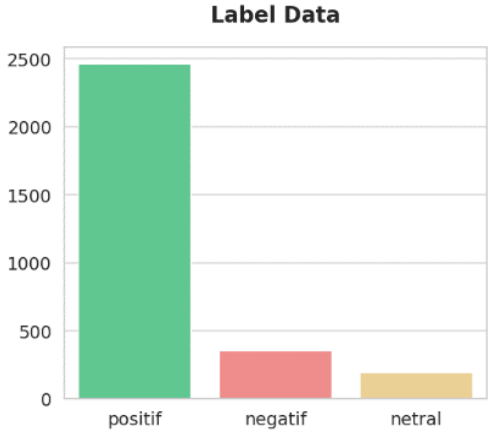
Tokenizing	*Normalization*	*Stopwords*	*Stemming*
['good', 'semakin', 'banyak', 'permainan', 'sip']	['bagus', 'semakin', 'banyak', 'permainan', 'mantap']	['bagus', 'banyak', 'permainan', 'mantap']	['bagus', 'banyak', 'main', 'mantap']
['wisatanya', 'bersih', 'rekomendasi', 'berkunjung', 'pada', 'saat', 'pagi', 'dan', 'sore']	['wisatanya', 'bersih', 'rekomendasi', 'berkunjung', 'pada', 'saat', 'pagi', 'dan', 'sore']	['bersih', 'rekomendasi', 'berkunjung', 'pagi', 'sore']	['bersih', 'rekomendasi', 'kunjung', 'pagi', 'sore']
['menjadi', 'alternatif', 'eduwisata', 'yg', 'ada', 'gresik', 'semoga', 'tambah', 'baik']	['menjadi', 'alternatif', 'eduwisata', 'yang', 'ada', 'gresik', 'semoga', 'tambah', 'baik']	['alternatif', 'eduwisata']	['alternatif', 'eduwisata']
['biasa', 'sj', 'panas', 'pengelolaan', 'ny', 'krg']	['biasa', 'saja', 'panas', 'pengelolaan', 'ny', 'kurang']	['panas', 'pengelolaan', 'kurang']	['panas', 'kelola', 'kurang']

Proses pelabelan dilakukan secara otomatis dengan pustaka *TextBlob* yang ditambah dengan kamus *lexicon* bahasa Indonesia. Hasil pelabelan seperti pada Tabel 8.

Tabel 8. Hasil Pelabelan

Setelah *pre-processing*	Polaritas	Label
bagus banyak main mantap	0,60000	positif
bersih rekomendasi kunjung pagi sore	0,20000	positif
alternatif eduwisata	0	netral
panas kelola kurang	-0,40000	negatif

Hasil pelabelan seluruh data yang telah selesai sebanyak 3012 data. Berikut rincian hasil persebaran label pada Gambar 2.



Gambar 2. Hasil Pelabelan Data

Dari seluruh data, mayoritas memiliki label positif. Tercatat ada 2462 data yang berlabel positif, 356 data dengan label negatif, dan 194 data yang memiliki label netral. Dari hasil pada Gambar 2, dilakukan validasi dengan pengelola wisata untuk melihat kecocokan label dari mesin. Dari teknik *judgement sampling* sebanyak 355 data digunakan sebagai sampel untuk dicek sentimen oleh 3 pengelola wisata yang berbeda pada objek wisata Kabupaten Gresik. Hasil validasi seperti pada Tabel 9.

Tabel 9. Validasi Pelabelan

Ulasan	Pengelola			Label *TextBlob*	M/U
	I	II	III		
Semakin indah, smg bsk makin bertambah wahana anak, mainan yg *free*,	1	0	1	1	M
Waktu kesana kebetulan pas tutup karena pandemi. Tdk bisa keliling melihat-lihat.	0	0	2	2	U
Tempatnya bagus tp syg bnyk yg g kerawat kyk kolam.e air.e kering	2	0	2	2	M
Tempatnya dekat dng makam	0	0	2	0	M
Tempatnya lumayan bagus,kasih wahana lagi kalau bisa..😁🙏 …	1	0	1	1	M

Keterangan angka 0 berarti netral, 1 berarti positif, dan 2 berarti negatif. Label M yaitu *match* atau cocok dan U berarti *unmatch* atau tidak cocok. Hasil keseluruhan validasi terdapat pada Tabel 10, yang memuat prediksi dari *TextBlob* dan pengelola. Secara akurasi mendapatkan nilai 81% yang menandakan bahwa label yang dibuat secara otomatis oleh mesin masih tergolong sesuai.

Tabel 10. Hasil Validasi

		Pengelola Wisata		
		Positif	**Negatif**	**Netral**
Label *TextBlob*	**Positif**	262	7	15
	Negatif	15	17	16
	Netral	13	3	7

Secara akurasi mendapatkan nilai 81% yang menandakan bahwa label yang dibuat secara otomatis oleh mesin masih tergolong sesuai.

$$\text{Akurasi} = \frac{262 + 17 + 7}{262 + 17 + 7 + 7 + 15 + 15 + 16 + 13 + 3} = \frac{286}{355} = 0.81$$

3.2. Feature Extraction

Data yang setelah melewati tahap *pre-processing* di vektorisasi menggunakan model *word embedding fasttext*. Setiap *vocabulary* memiliki nilai vektor. Nilai tersebut dihitung kembali rata-ratanya sesuai kata pada setiap baris data. ApabiIa terdapat kata yang tidak terdeteksi atau tidak termasuk dalam *vocabulary* model, nilai vektor dianggap 0. Hasil vektor setelah direduksi seperti pada Tabel 11.

Tabel 11. Hasil Vektorisasi Model Dan *Dimensionality Reduction*

Preprocess	**Nilai Vektor**	**LDA**
bagus malam	[-0.012589 -0.04042 -0.022949 -0.081458 0.007692 -0.058139...]	[-0.08218208 -0.9504243]
pandang tangga bagus air	[-0.01083 -0.042977 -0.024336 -0.083083 0.00147 -0.051627...]	[0.09743813 0.6172086]
tidak sesuai alamat	[-0.018289 -0.037894 -0.025132 -0.086528 0.003465 -0.043483....]	[-2.07791728 2.66395544]

Dimensi vektor pada setiap kalimat sebanyak 300. Pada tahap *dimensionality reduction,* ukuran dimensi akan direduksi dan menyisihkan 2 dimensi. Proses reduksi menggunakan *Linear Discriminant Analysis* pada pustaka *Sklearn.*

3.3. Analisis Sentimen

Proses berlanjut ke *cross validation.* Pengujian dilakukan dengan nilai k 10. Data akan melakukan pelatian setiap *fold* dari nilai k yang diberikan. Pada tahap ini, penentuan parameter dilakukan dahulu dari *GridSearchCV*. Pada Tabel 12, hasil dari proses *Hyperparameter Tuning* menghasilkanparameter terbaik yang diperoleh dari model.

Tabel 12. Hasil *Hyperparameter Tuning*

Parameter	**Model**	
	SVM	**SVM+LDA**
C	100	100
Kernel	RBF	RBF
Gamma	10	1

Parameter dari tiap model selanjutnya dilakukan pengujian *Cross Validation Score* sesuai yang ditentukan sebelumnya dengan *scoring* berdasarkan *F1-Macro.* Setiap *fold* memiliki hasil yang berbeda, selengkapnya ditampilkan pada Tabel 13.

Tabel 13. Hasil *K-Fold Cross Validation*

Fold	***Score***	
	SVM	**SVM+LDA**
1	0,44977188	0,74062619
2	0,47902639	0,77060932
3	0,52105509	0,74158551
4	0,46044487	0,72047242
5	0,61169066	0,81278233
6	0,4821534	0,80413089
7	0,47967153	0,81876484
8	0,47214486	0,88559671
9	0,55595188	0,80861678
10	0,52051282	0,7818057
Mean	**0,50324234**	**0,78849907**

Berdasarkan hasil pengujian *cross validation*, penerapan LDA dapat meningkatkan evaluasi *F1-score*. Hasil rata-rat skor model SVM dengan LDA mencapai angka 0,78 terpaut jauh di bawahnya model SVM tanpa LDA dengan skor 0,50. Berdasarkan hasil ini, model SVM dengan LDA terlihat lebih unggul. Selanjutnya proses *modeling* yang dilakukan dengan rasio pembagian data 80:20, 2409 data latih dan 603 data uji dengan parameter dari hasil *hyperparameter tuning.* Evaluasi model berdasarkan dari *confusion matrix.* Hasil dari model SVM terdapat pada Tabel 14.

Tabel 14. Evaluasi Model SVM

	Label		
	Netral	**Positif**	**Negatif**
Precision	1,00	0.86	0.73
Recall	0.15	0.99	0.27
F1-Score	0.27	0.27	0.39
Accuracy	0.85		

Pada model SVM tanpa LDA diperoleh akurasi 85%. Pada setiap kelas menghasilkan nilai evaluasi yang berbeda. *Precision* label netral memperoleh angka 100%, hasil ini tidak sesuai dengan keadaan persebaran label data pada Gambar 2, bahwa label netral dalam kondisi minoritas. Pada hasil *recall,* nilai yang dihasilkan masih rendah pada label minoritas, hasil ini berpengaruh terhadap nilai *f1-score* pada label tersebut.

Selanjutnya pada model SVM dengan LDA menghasilkan evaluasi yang berbeda. Sesuai dengan proses *cross validatioon* sebelumnya pada Tabel 13, pada model ini memang menunjukkan hasil yang lebih baik. Hasil evaluasi SVM dengan LDA pada Tabel 15.

Tabel 15. Evaluasi Model SVM+LDA

	Label		
	Netral	**Positif**	**Negatif**
Precision	0.56	0.92	0.62
Recall	0.38	0.95	0.56
F1-Score	0.45	0.94	0.59
Accuracy	0.87		

Hasil akurasi model SVM dengan LDA sebesar 87%, secara akurasi lebih unggul dari model sebelumnya. Nilai evaluasi pada tiap model sudah lebih optimal walaupun pada label netral, nilai *recall* masih rendah. Hasil tersebut mempengaruhi nilai *f1-score* dari label yang memiliki nilai rendah. Berikut hasil perbandingan metrik evaluasi tiap model pada Tabel 16.

Tabel 16. Perbandingan Hasil Evaluasi

	Mode1	
	SVM	**SVM+LDA**
Precision	86%	70%
Recall	47%	63%
F1-Score	53%	66%
Accuracy	85%	87%

Berdasarkan perbandingan setiap metrik evaluasi model SVM dengan reduksi dimensi LDA lebih unggul daripada model SVM tanpa reduksi dimensi dari LDA. Perhitungan nilai metrik lainnya dari *precision, recall,* dan *f1-score* diambil dari sisi *macro average,* dari hasil setiap label pada hasil evaluasi sebelumnya.

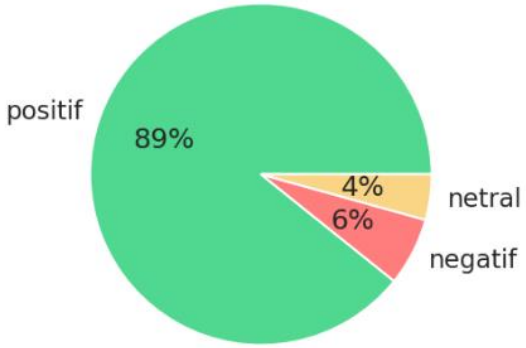
Dari hasil *precision* secara kalkulasi model SVM saja tanpa reduksi dimensi dari LDA lebih tinggi dengan 86% berbeda 16% dari model SVM dengan reduksi dimensi LDA yaitu 70%. Sebelumnya pada label netral pada model SVM menghasilkan nilai *precision* 100% sedangkan label tersebut dalam kondisi minoritas, berdasarkan hal tersebut, model SVM mengalami *overfit,* walaupun hasil model SVM dengan LDA hanya 70%, tetapi sesuai dengan kondisi data yang diberikan.

Hasil perbandingan nilai *recall*, model SVM dengan reduksi dimensi LDA lebih unggul dengan 63% daripada model SVM tanpa LDA dengan 47%. Pada perhitungan nilai *f1-score*, model SVM dengan reduksi dimensi LDA memperoleh nilai 66%. Sedangkan pada model SVM tanpa reduksi dimensi LDA, nilai *recall*-nya rendah, berpengaruh terhadap hasil *f1-score* 53%.

Berdasarkan perbandingan yang telah dilakukan, model SVM dengan reduksi dimensi dari LDA memiliki hasil yang lebih baik. Selanjutnya model tersebut digunakan dalam memprediksi sentimen pada ulasan objek wisata di Kabupaten Gresik. Data untuk prediksi difokuskan hanya data pada tahun 2023, karena sesuai dengan indikasi penurunan pengunjung pada tahun tersebut dan diperoleh data sebanyak 534 ulasan..

Proses prediksi sama dengan tahapan sebelumnya, diawali eksplorasi data untuk mengetahui kolom kosong dan duplikat data. Berikutnya *preprocessing,* dengan tahapan

dan proses yang sama. Hasil data setelah tahap *preprocessing* menjadi 511 ulasan. Tahap berikutnya pelaksanaan klasifikasi sentimen dengan model SVM+LDA.



Gambar 3. Hasil Prediksi Sentimen

Hasil klasifikasi sentimen pada Gambar 3, menunjukkan bahwa sentimen dari ulasan pengunjung wisata di Kabupaten Gresik mengarah ke argumen positif. Jumlah rincian sentimen yaitu positif sebanyak 456 ulasan, negatif sebanyak 33 ulasan, dan netral sebanyak 22 ulasan.

Gambar 4. *Wordcloud* Sentimen Positif

Pada Gambar 4, frekuensi kata pada sentimen positif, yang paling sering digunakan, “bagus”, “bersih”, “wahana”, “main” “nyaman”, dan “keluarga”. Kata-kata ini mengindikasikan keadaan objek wisata di Kabupaten Gresik yang memang menurutnya beberapa pengunjung bersih, bagus dan nyaman. Kata lainnya, “wahana”, “main” dan “keluarga”. menunjukkan bahwa sebagian objek wisata cocok untuk liburan bersama keluarga dengan setiap wahana yang ada untuk bermain bagi anak-anak.

Gambar 5. *Wordcloud* Sentimen Negatif

Pada sentimen negatif, beberapa kata dengan frekuensi tinggi yaitu, “panas", "tiket", "parkir", dan "mahal". Gambar 5 menunjukkan kata “panas” menjadi kata yang paling banyak kemunculannya. Secara umum, kondisi cuaca di Kabupaten Gresik memang panas. Terdapat Kata ‘tiket” dan “mahal” yang dapat diartikan bahwa harga tiket pada salah satu objek wisata di Kabupaten Gresik dinilai masih mahal oleh sebagian pengunjung. Kata selanjutnya “parkir” kemungkinan kondisi lahan yang kurang luas atau penempatan posisi lokasi parkir yang kurang pas.

Berdasarkan hasil analisis sentimen ulasan objek wisata di Kabupaten Gresik lebih mengarah ke sentimen positif dibandingkan sentimen lainnya. Hasil ini dapat menjadi solusi pengelola wisata dalam meningkatkan fasilitas maupun infrastruktur pada objek wisatanya masing-masing. Walaupun dari hasil sentimen pengunjung terlihat positif, peningkatan kualitas tetap selalu diperhatikan pengunjung dapat merasa nyaman dan meningkat dari sebelumnya.

4. KESIMPULAN

Hasil dari klasifikasi sentimen dari 511 ulasan, diperoleh 456 sentimen positif, 33 sentimen negatif, dan 22 sentimen netral. Model SVM dengan LDA dapat digunakan dalam klasifikasi sentimen dengan baik. Berdasarkan hasil tersebut perspektif para pengunjung terhadap objek wisata di Kabupaten Gresik mengarah pada sentimen positif.

Model SVM dengan menerapkan LDA atau SVM-LDA dapat menghasilkan metrik evaluasi yang lebih unggul dibandingkan dari model SVM. Model SVM-LDA memperoleh nilai *F1-score* 66%, unggul 11 % daripada model SVM (tidak dengan LDA) yang bernilai *F1-score* 55%. Model SVM-LDA mengalami *overfit* yang mengakibatkan penurunan nilai evaluasi, sehingga selisih 11% antara kedua metode dengan satuan *F1-Score* menunjukkan bahwa pengurangan vektorisasi pada SVM sangat efektif untuk meningkatkan kualitas klasifikasi sentimen dengan SVM.

5. SARAN

Word embedding model berfungsi untuk memahami hubungan konteks antar kata. Kualitas model sangat dipengaruhi oleh ukuran dan kualitas data. Penerapan model *word embedding* yang sudah ada atau training model baru dengan korpus lebih besar menerapkan metode vektorisasi lainnya, dengan GloVe atau *Word2Vec*.

Memberikan rekomendasi peningkatan metode analisis dengan metode *deep learning* dengan LDA agar analisis lebih akurat, dan penanganan terhadap kondisi *imbalance* data berikutnya terutama pada data teks dapat melakukan penambahan metode *text augmentation*.

UCAPAN TERIMA KASIH

Terima kasih kepada para pengelola wisata yang telah memberikan validasi dan informasi kepada peneliti, dan kepada penyedia korpus dari *Kaggle* yang telah mempublikasikan data tersebut sehingga dapat saya implementasikan pada penelitian ini.

DAFTAR PUSTAKA

Agustiningsih, K. K., Utami, E., & Alsyaibani, M. A. (2022). Sentiment Analysis of COVID-19 Vaccines in Indonesia on Twitter Using Pre-Trained and Self-Training Word Embeddings. *Jurnal Ilmu Komputer Dan Informasi*, *15*(1), 39–46. https://doi.org/10.21609/jiki.v15i1.1044

Anasta, D. (2023). *Sentiment Analysis Review of The Sayurbox App*.

Cahyanti, F. E., Adiwijaya, & Faraby, S. Al. (2020). On The Feature Extraction For Sentiment Analysis of Movie Reviews Based on SVM. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1–5. https://doi.org/10.1109/ICoICT49345.2020.9166397

ÇeliK, Ö., & Koç, B. C. (2021). Classification of Turkish News Texts with TF-IDF, Word2vec and Fasttext Vector Model Methods. *Deu Muhendislik Fakultesi Fen ve Muhendislik*, *23*(67), 121–127. https://doi.org/10.21205/deufmd.2021236710

Dakuwison. (2018). *Data Kunjungan Wisata Online*.

Diandra, D. (2022). *Analisis Sentimen Ulasan MyXL dengan SVM*.

Haq, F. U. (2020). Penggunaan Google Review Sebagai Penilaian Kepuasan Pengunjung Dalam Pariwisata. *Tornare*, *2*(1), 10. https://doi.org/10.24198/tornare.v2i1.25826

Herlawati, H., Handayanto, R. T., Atika, P. D., Khasanah, F. N., Yusuf, A. Y. P., & Septia, D. Y. (2021). Analisis Sentimen Pada Situs Google Review dengan Naïve Bayes dan Support Vector Machine. *Jurnal Komtika (Komputasi dan Informatika)*, *5*(2), 153–163. https://doi.org/10.31603/komtika.v5i2.6280

Hesay, I. K., Indiriati, & Adinugroho, S. (2021). Analisis Sentimen Ulasan Pengunjung Simpang Lima Gumul Kediri menggunakan Metode BM25 dan Neighbor-Weighted K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *5*(7), 3160–3169.

Lai, V., & Tan, C. (2019). *On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection*. 29–38. https://doi.org/10.1145/3287560.3287590

Lappeman, J., Clark, R., Evans, J., Sierra-Rubia, L., & Gordon, P. (2020). Studying social media sentiment using human validated analysis. *MethodsX*, *7*, 100867. https://doi.org/10.1016/j.mex.2020.100867

Larasati, L., Nabilla, S., & Haryanto, E. (2022). Sentiment Analysis untuk Review Destinasi Wisata Unggulan Gunung Kidul Menggunakan Metode Lexicon Dan

Pivot. *Indonesian Journal of Business Intelligence (IJUBI)*, *5*(2), 102. https://doi.org/10.21927/ijubi.v5i2.2604

Prasetyo, D. B., & Hidayatullah, A. F. (2020). Identifikasi Dual Sentimen Terhadap Ulasan Objek Wisata di Daerah Istimewa Yogyakarta. *AUTOMATA*, *1*(1).

Pratama, Y. T., Bachtiar, F. A., & Setiawan, N. Y. (2018). Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF Dan Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(12), 6244–6252.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, *8*, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

Santika, E. F. (2023). *Wisatawan Jawa Timur Mendominasi Arus Pariwisata Domestik 2022 | Databoks*.

Saputra, R. R. (2022). Daftar 10 Provinsi Terkaya di Indonesia, Nomor 7 Ditempati IKN Nusantara. Dalam *iNews.ID*.

Suryawan, I. W. B., Utami, N. W., & Fredlina, K. Q. (2023). Analisis Sentimen Review Wisatawan Pada Objek Wisata Ubud Menggunakan Algoritma Support Vector Machine. *Jurnal Informatika Teknologi dan Sains*, *5*(1), 133–140. https://doi.org/10.51401/jinteks.v5i1.2242

Utami, D. S., & Erfina, A. (2022). Analisis Sentimen Objek Wisata Bali Di Google Maps Menggunakan Algoritma Naive Bayes. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, *6*(1), 418–427.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224