SKRIPSI

PENGARUH RFE TERHADAP *LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE* PADA ANALISIS SENTIMEN HOTEL SHANGRI-LA SURABAYA

FAKHRI MAULANA HERZA

NPM 20081010039

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S. Si. MT.

M. Muharrom AL Haromainy, S.Kom., M.Kom.

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2024

SKRIPSI

PENGARUH RFE TERHADAP *LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE* PADA ANALISIS SENTIMEN HOTEL SHANGRI-LA SURABAYA

FAKHRI MAULANA HERZA

NPM 20081010039

DOSEN PEMBIMBING

Dr. Basuki Rahmat, S. Si. MT.

M. Muharrom AL Haromainy, S.Kom., M.Kom.

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI

UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI INFORMATIKA

SURABAYA

2024

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

PENGARUH RFE TERHADAP *LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE* PADA ANALISIS SENTIMEN HOTEL SHANGRI-LA SURABAYA

Oleh:
FAKHRI MAULANA HERZA
NPM. 20081010039

Telah dipertahankan dihadapan dan diterima oleh Tim Penguji Skripsi Prodi Informatika Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jawa Timur Pada tanggal 2 September 2024

Menyetujui

Dr. Basuki Rahmat, S. Si. MT. ..... **(Pembimbing I)**
NIP. 19690723 2021211 002

M. Muharrom AL Haromainy, S.Kom., M.Kom. ..... **(Pembimbing II)**
NIP. 19950601 202203 1 006

Dr. I Gede Susrama Mas Diyasa, ST., MT. IPU. ..... **(Ketua Penguji)**
NIP. 19700619 2021211 009

Retno Mumpuni, S.Kom., M.Sc. ..... **(Anggota Penguji)**
NPT. 172198 70 716054

Mengetahui,
Dekan Fakultas Ilmu Komputer

Prof. Dr. Ir. Novirina Hendrasarie, MT.
NIP. 19681126 199403 2 001

Dipindai dengan CamScanner

LEMBAR PERSETUJUAN

PENGARUH RFE TERHADAP *LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE* PADA ANALISIS SENTIMEN HOTEL SHANGRI-LA SURABAYA

Oleh:

FAKHRI MAULANA HERZA

NPM. 20081010039

Menyetujui,

Koordinator Program Studi Informatika

Fakultas Ilmu Komputer

Fetty Tri Anggraeny, S.Kom., M.Kom.

NIP. 19820211 2021212 005

Dipindai dengan CamScanner

Halaman ini sengaja dikosongkan

SURAT PERNYATAAN ORISINALITAS

Yang bertandatangan di bawah ini:

Nama Mahasiswa : Fakhri Maulana Herza

Program Studi : Informatika

Dosen Pembimbing : Dr. Basuki Rahmat, S.Si. MT.

dengan ini menyatakan bahwa isi sebagian maupun keseluruhan disertasi dengan judul:

PENGARUH RFE TERHADAP *LOGISTIC REGRESSION* DAN *SUPPORT VECTOR MACHINE* PADA ANALISIS SENTIMEN HOTEL SHANGRI-LA SURABAYA

adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diizinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri. Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, 2 September 2024

Yang Membuat Pernyataan,

FAKHRI MAULANA HERZA

NPM. 20081010039

Dipindai dengan CamScanner

Halaman ini sengaja dikosongkan

ABSTRAK

Nama Mahasiswa / NPM	: Fakhri Maulana Herza / 20081010039
Judul Skripsi	: Pengaruh RFE Terhadap *Logistic Regression* Dan *Support Vector Machine* Pada Analisis Sentimen Hotel Shangri-La Surabaya
Dosen Pembimbing	: 1. Dr. Basuki Rahmat, S.Si. MT. 2. M. Muharrom AL Haromainy, S.Kom., M.Kom

Analisis sentimen adalah salah satu sarana penting dalam industri pariwisata untuk memahami respon dan pengalaman tamu terhadap layanan hotel. Ulasan para tamu sebelumnya berperan krusial dalam membentuk persepsi calon tamu terhadap kualitas fasilitas dan daya tarik hotel yang akan dikunjungi. Namun, salah satu tantangan utama dalam analisis sentimen adalah memilih fitur yang paling relevan untuk meningkatkan kinerja model prediksi. Banyak ulasan mengandung informasi yang beragam, dan tidak semua kata atau fitur memiliki kontribusi yang signifikan dalam membedakan antara sentimen positif dan negatif.

Dalam konteks ini, metode *Recursive Feature Elimination* (RFE) berpotensi untuk mengoptimalkan pemilihan fitur dengan mengeliminasi fitur yang kurang informatif, sehingga diharapkan mampu meningkatkan akurasi model *Logistic Regression* dan *Support Vector Machine* (SVM) dalam analisis sentimen. Oleh karena itu, penelitian ini fokus pada pengaruh penerapan RFE terhadap kinerja kedua model tersebut, khususnya dalam analisis sentimen ulasan tamu di Hotel Shangri-La Surabaya. Data yang digunakan dalam penelitian ini berjumlah 3719 ulasan.

Hasil pengujian menunjukkan bahwa pada model *Logistic Regression* yang menggunakan RFE, terdapat peningkatan performa yang signifikan dalam nilai presisi, sensitivitas, *F1 Score*, dan akurasi, dengan rata-rata peningkatan sebesar 9%. Bahkan, untuk model *Support Vector Machine* yang menggunakan RFE, peningkatan performa lebih signifikan, dengan rata-rata peningkatan sebesar 14%. Temuan ini menunjukkan bahwa penerapan RFE secara efektif dapat meningkatkan kualitas prediksi pada kedua model dalam konteks analisis sentimen ulasan hotel.

Kata kunci : Analisis Sentimen, Sektor Pariwisata, *Google Maps*, *Recursive Feature Elimination*, *Support Vector Machine*, *Logistic Regression*.

Halaman ini sengaja dikosongkan

ABSTRACT

Student Name / NPM	: Fakhri Maulana Herza / 20081010039
Thesis Title	: The Influence Of RFE On Logistic Regression And Support Vector Machine In Sentiment Analysis Of The Shangri-La Hotel Surabaya
Advisor	: 1. Dr. Basuki Rahmat, S.Si. MT. 2. M. Muharrom AL Haromainy, S.Kom., M.Kom

Sentiment analysis is one of the essential tools in the tourism industry for understanding guest responses and experiences regarding hotel services. Previous guest reviews play a crucial role in shaping the perceptions of potential guests about the quality of facilities and the attractiveness of the hotel they plan to visit. However, one of the main challenges in sentiment analysis is selecting the most relevant features to improve model prediction performance. Many reviews contain diverse information, and not all words or features significantly contribute to distinguishing between positive and negative sentiment.

In this context, the Recursive Feature Elimination (RFE) method has the potential to optimize feature selection by eliminating less informative features, thus expected to improve the accuracy of Logistic Regression and Support Vector Machine (SVM) models in sentiment analysis. Therefore, this study focuses on the impact of RFE application on the performance of both models, particularly in analyzing guest reviews at the Shangri-La Hotel Surabaya. The data used in this study consists of 3719 reviews.

The test results show that in the Logistic Regression model using RFE, there was a significant improvement in performance in terms of precision, sensitivity, F1 Score, and accuracy, with an average increase of 9%. Moreover, for the Support Vector Machine model using RFE, the performance improvement was even more significant, with an average increase of 14%. These findings indicate that the application of RFE can effectively enhance the predictive quality of both models in the context of hotel review sentiment analysis.

Kata kunci : *Sentiment Analysis, Tourism Sector, Google Maps, Recursive Feature Elimination, Support Vector Machine, Logistic Regression.*

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas segala rahmat, hidayah dan karunia-Nya kepada penulis sehingga skripsi dengan judul **"Pengaruh RFE Terhadap *Logistic Regression* Dan *Support Vector Machine* Pada Analisis Sentimen Hotel Shangri-La Surabaya"** dapat terselesaikan dengan baik.

Oleh karena itu, saya ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada semua pihak yang telah berkontribusi.

1. Prof. Dr. Ir. Akhmad Fauzi, M.MT selaku Rektor Universitas Pembangunan Nasional "Veteran" Jawa Timur.
2. Prof. Dr. Novirina Hendrasarie, S.T, M.T. Selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional "Veteran" Jawa Timur.
3. Ibu Fetty Tri Anggraeny, S.Kom, M.Kom selaku Ketua Program Studi Teknik Informatika Universitas Pembangunan Nasional "Veteran" Jawa Timur.
4. Bapak Dr. Basuki Rahmat, S.Si. MT selaku dosen pembimbing 1 dan bapak M. Muharrom Al Haromainy, S.Kom., M.Kom selaku dosen pembimbing 2 yang telah membimbing serta memberikan arahan dalam menyusun pengerjaan tugas akhir ini.
5. Orang Tua tercinta saya yang telah memberikan doa, dukungan, dan motivasi yang diberikan kepada penulis.

Saya menyadari bahwa laporan ini masih memiliki beberapa kekurangan. Oleh karena itu, saya sangat mengharapkan saran dan kritik yang konstruktif demi penyempurnaan laporan ini. Akhir kata, dengan penuh harap akan Ridho dari Allah SWT, semoga laporan tugas akhir ini dapat bermanfaat bagi kita semua. Aamiin.

Surabaya, 2 September 2024

Penulis

Halaman ini sengaja dikosongkan

DAFTAR ISI

LEMBAR PENGESAHAN iii
SURAT PERNYATAAN ORISINALITAS v
ABSTRAK vii
KATA PENGANTAR xi
DAFTAR ISI xiii
DAFTAR GAMBAR xv
DAFTAR TABEL xvii
DAFTAR NOTASI xix
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan 4
1.4 Manfaat 4
1.5 Batasan Masalah 5
BAB II TINJAUAN PUSTAKA 7
2.1 Penelitian Sebelumnya 7
2.2 Shangri-La 8
2.3 Analisis Sentimen 9
2.4 Preprocessing 9
2.4.1 Case-folding 10
2.4.2 Cleaning 10
2.4.3 Tokenizing 10
2.4.4 Stopword Removal 10
2.4.5 Stemming 10
2.5 TF-IDF 10
2.6 Recursive Feature Elimination (RFE) 12
2.7 Support Vector Machine (SVM) 12
2.8 Logistic Regression 14
2.9 Confusion Matrix 16
BAB III METODOLOGI 19
3.1 Tahapan Penelitian 19
3.2 Studi Literatur 19
3.3 Analisa Dan Desain 20

3.4 Pengambilan Dan *Labeling* Data 21
3.5 Preprocessing Data 23
3.5.1 Case Folding 24
3.5.2 Cleaning 24
3.5.3 Tokenizing 24
3.5.4 Stopword Removal 26
3.5.5 Stemming 27
3.6 TF-IDF 28
3.7 Recursive Feature Elimination (RFE) 29
3.8 Support Vector Machine (SVM) 31
3.9 Logistic Regression 38
3.10 Skenario Pengujian 41
BAB IV HASIL DAN PEMBAHASAN 43
4.1 Akusisi dan *Labeling* Data 43
4.2 Preprocessing data 45
4.2.1 Case Folding 47
4.2.2 Cleaning 48
4.2.3 Tokenizing 49
4.2.4 Stopword Removal 50
4.2.5 Stemming 51
4.3 TF-IDF 52
4.4 Splitting Data 54
4.5 Recursive Feature Elimination (RFE) 55
4.6 Support Vector Machine (SVM) 57
4.6.1 SVM menggunakan RFE 58
4.6.2 SVM tidak menggunakan RFE 64
4.7 Logistic Regression 70
4.7.1 *Logistic Regression* menggunakan RFE 71
4.7.2 *Logistic Regression* tidak menggunakan RFE 77
4.8 Hasil Evaluasi Model 83
4.9 Analisis Sentimen Aspek 88
BAB IV KESIMPULAN DAN SARAN 91
5.1 Kesimpulan 91
5.2 Saran 92
DAFTAR PUSTAKA 93

DAFTAR GAMBAR

Gambar 2. 1 Garis *Hyperlane* SVM.....13
Gambar 2. 2 Klasifikasi SVM 13
Gambar 2. 3 Contoh Rumus *Logistic Regression* 15
Gambar 2. 4 Contoh Bentuk *Logistic Regression*.....15
Gambar 3. 1 Flowchart Tahapan Penelitian 19
Gambar 3. 2 Flowchart Rancangan Sistem.....20
Gambar 3. 3 Flowchart Pengambilan Dan *Labeling* Data 21
Gambar 3. 4 *Scrapping* Dataset.....22
Gambar 3. 5 Contoh *Labelling* Dataset.....23
Gambar 3. 6 Flowchart *Preprocessing* Data 23
Gambar 3. 7 Flowchart RFE.....29
Gambar 3. 8 Flowchart SVM.....32
Gambar 3. 9 Flowchart *Logistic Regression* 38
Gambar 4. 1 Hasil *Scrapping* data 43
Gambar 4. 2 Hasil Pelabelan Sentimen Positif dan Negatif.....44
Gambar 4. 3 Hasil Upload *File* Dataset Pada *Google Collab* 46
Gambar 4. 4 Dataset Hasil *Upload* 47
Gambar 4. 5 Hasil Proses *Case-folding* 48
Gambar 4. 6 Hasil Proses *Cleaning* 49
Gambar 4. 7 Hasil Proses *Tokenizing*.....50
Gambar 4. 8 Hasil Proses *Stopword Removal* 51
Gambar 4. 9 Hasil Proses *Stemming*.....52
Gambar 4. 10 Hasil Array *Label Encoding*.....53
Gambar 4. 11 Hasil Proses TF-IDF 54
Gambar 4. 12 Data Sebelum Diproses RFE.....56
Gambar 4. 13 Data Setelah Diproses RFE.....57
Gambar 4. 14 *Confusion Matrix* 90% Data Train Dan 10% Data Test.....59
Gambar 4. 15 *Confusion Matrix* 80% Data Train Dan 20% Data Test.....61
Gambar 4. 16 *Confusion Matrix* 70% Data Train Dan 30% Data Test.....63
Gambar 4. 17 *Confusion Matrix* 90% Data Train Dan 10% Data Test.....65
Gambar 4. 18 *Confusion Matrix* 80% Data Train Dan 20% Data Test.....67
Gambar 4. 19 *Confusion Matrix* 70% Data Train Dan 30% Data Test.....69
Gambar 4. 20 *Confusion Matrix* 90% Data Train Dan 10% Data Test.....72
Gambar 4. 21 *Confusion Matrix* 80% Data Train Dan 20% Data Test.....74
Gambar 4. 22 *Confusion Matrix* 70% Data Train Dan 30% Data Test.....76
Gambar 4. 23 *Confusion Matrix* 90% Data Train Dan 10% Data Test.....78
Gambar 4. 24 *Confusion Matrix* 80% Data Train Dan 20% Data Test.....80
Gambar 4. 25 *Confusion Matrix* 70% Data Train Dan 30% Data Test.....82
Gambar 4. 26 Diagram Batang Akurasi 85
Gambar 4. 27 Diagram Batang Presisi.....86
Gambar 4. 28 Diagram Batang Sensitifitas 87
Gambar 4. 29 Diagram Batang *F1 Score*.....88
Gambar 4. 30 Analisis Sentimen Aspek.....89

Halaman ini sengaja dikosongkan

DAFTAR TABEL

Tabel 3. 1 Contoh *Case Folding* 24
Tabel 3. 2 Contoh *Cleaning* 24
Tabel 3. 3 Contoh *Tokenizing* 24
Tabel 3. 4 Contoh *Stopword Removal* 26
Tabel 3. 5 Contoh *Stemming* 27
Tabel 3. 6 Contoh Bobot Setiap Fitur 30
Tabel 3. 7 Contoh Kelas Sentimen 32
Tabel 3. 8 Contoh Ektraksi Fitur Pada Dataset 33
Tabel 3. 9 Contoh Pasangan Kernel 34
Tabel 3. 10 Contoh hasil Perhitungan Kernelisasi 34
Tabel 3. 11 Contoh Hasil Perhitungan Matriks Hessian 35
Tabel 3. 12 Data Kernel 35
Tabel 3. 13 Matriks Kernel 37
Tabel 3. 14 Contoh Dataset 38
Tabel 3. 15 Skenario Pengujian 41
Tabel 4. 1 Kode Program *Import Library Preprocessing* Data 45
Tabel 4. 2 Kode Program *Input File* Dataset 46
Tabel 4. 3 Kode Program *Case Folding* 47
Tabel 4. 4 Kode Program *Cleaning* 48
Tabel 4. 5 kode program *Tokenizing* 49
Tabel 4. 6 Kode Program *Stopword Removal* 50
Tabel 4. 7 Kode Program *Stemming* 51
Tabel 4. 8 Kode Program *Label Encoding* 52
Tabel 4. 9 Kode Program TF-IDF 53
Tabel 4. 10 Kode Program *Splitting Data* 54
Tabel 4. 11 Metode Optimasi RFE 55
Tabel 4. 12 Kode Program *Support Vector Machine* 57
Tabel 4. 13 Hasil Evaluasi Model 60
Tabel 4. 14 Hasil Evaluasi Model 62
Tabel 4. 15 Hasil Evaluasi Model 64
Tabel 4. 16 Hasil Evaluasi Model 66
Tabel 4. 17 Hasil Evaluasi Model 68
Tabel 4. 18 Hasil Evaluasi Model 70
Tabel 4. 19 Kode Program *Logistic Regression* 70
Tabel 4. 20 Hasil Evaluasi Model 73
Tabel 4. 21 Hasil Evaluasi Model 75
Tabel 4. 22 Hasil Evaluasi Model 77
Tabel 4. 23 Hasil Evaluasi Model 79
Tabel 4. 24 Hasil Evaluasi Model 81
Tabel 4. 25 Hasil Evaluasi Model 83
Tabel 4. 26 Hasil Evaluasi Seluruh Model 83

Halaman ini sengaja dikosongkan

DAFTAR NOTASI

TF	: Frekuensi kemunculan kata pada suatu dokumen d.
td	: Jumlah kemunculan kata t dalam dokumen d.
d	: Total jumlah kata dalam dokumen d.
lg	: logaritma natural
D	: jumlah semua dokumen.
df	: jumlah *term*/kata dari dokumen.
W	: bobot
TF	: Nilai TF
IDF	: Nilai IDF
D_{ij}	: Elemen matriks data ke – ij
y_i	: Kelas atau label data ke – i
y_j	: Kelas atau label data ke – j
λ^2	: Turunan batas teoritis
E_i	: Nilai error data ke – I
γ	: Tingkat pembelajaran
$Max_i D_{ij}$	: Nilai maksimum diagonal matriks hessian
c	: 1 (Konstanta untuk perhitungan kernel)
λ	: 0.5
γ	: 0.001 (*learning rate*)
C	: 1 (*Variabel slack*)
ε	: 0.0001 (epilson)

Halaman ini sengaja dikosongkan