

**ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR
MUTUAL INFORMATION, N-GRAM DAN CHI-SQUARE
MENGUNAKAN METODE SVM PADA IMDB REVIEW**

SKRIPSI

Sevira Chairunnisa Nandaresta : 20200050009

Fira Elfariani 20200050079

Mirna Sari Gusti R 20200050029



**FAKULTAS TEKNIK KOMPUTER DAN DESAIN
PROGRAM STUDI SISTEM INFORMASI
UNIVERSITAS NUSAPUTRA
JUNI 2024**

**ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR
MUTUAL INFORMATION, N-GRAM DAN CHI-SQUARE
MENGUNAKAN METODE SVM PADA IMDB REVIEW**

SKRIPSI

*Diajukan Untuk Memenuhi Salah Satu Syarat Dalam Menempuh
Gelar Sarjana Komputer*

Sevira Chairunnisa Nandaresta 20200050009

Fira Elfariani 20200050079

Mirna Sari Gusti R 20200050029



**PROGRAM STUDI SISTEM INFROMASI
FAKULTAS TEKNIK KOMPUTER DAN DESAIN
UNIVERSITAS NUSAPUTRA
SUKABUMI
2024**

PERNYATAAN PENULIS

JUDUL : ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR
MUTUAL INFORMATION, N-GRAM DAN CHI-SQUARE
MENGGUNAKAN METODE SVM PADA IMDB REVIEW

NAMA : Sevira Chairunnisa Nandaresta NIM : 20200050009

NAMA : Fira Elfariani NIM : 20200050079

NAMA : Mirna Sari Gusti R NIM : 20200050029

“Kami menyatakan dan bertanggung jawab dengan sepenuhnya bahwa Skripsi ini adalah hasil karya kami kecuali cuplikan dan ringkasan yang masing-masing telah kami jelaskan sumbernya. Jika pada waktu selanjutnya ada pihak lain yang mengklaim bahwa Skripsi ini sebagai karyanya, yang disertai dengan bukti-bukti yang cukup, maka kami bersedia untuk dibatalkan gelar Sarjana Komputer kami beserta segala hak dan kewajiban yang melekat pada gelar tersebut”.

Sukabumi, 11 Juni 2024

Penulis

Penulis

Penulis

Sevira Chairunnisa N
20200050009

Fira Elfariani
20200050079

Mirna Sari Gusti R
20200050029

PENGESAHAN SKRIPSI

JUDUL : ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR
MUTUAL INFORMATION, N-GRAM DAN CHI-SQUARE
MENGUNAKAN METODE SVM PADA IMDB
REVIEW

NAMA PENULIS 1 : Sevira Chairunnisa Nandaresta NIM : 20200050009

NAMA PENULIS 2 : Fira Elfariani NIM : 20200050079

NAMA PENULIS 3 : Mirna Sari Gusti R NIM : 20200050029

Skripsi ini telah diajukan dan dipertahankan di depan Dewan Penguji pada
Sidang Skripsi tanggal 11 Juni 2024 Menurut pandangan kami, Skripsi ini memadai
dari segi kualitas untuk tujuan penganugerahan gelar Sarjana Komputer (S.Kom)

Sukabumi, 11 Juni 2024

Dosen Pembimbing I

Dosen Pembimbing II

Adhitia Erfina, S.T., M.Kom
NIDN. 0417049102

Dede Sukmawan, S.Kom, M.Kom
NIDN. 0405119501

Ketua Penguji

**Ketua Program Studi
Sistem Informasi**

Sudin Saepudin M.Kom
NIDN. 0414088608

Adhitia Erfina, S.T., M.Kom
NIDN. 0417049102

Plh. Dekan Fakultas Teknik Komputer dan Desain

Ir. Paikun, S.T, M.T, IPM
NIDN : 0402037401

ABSTRACT

Sentiment analysis is the computational process of extracting, finding, or assessing opinions or sentiments in text, such as reviews, social media posts, news articles, or other text types. In sentiment analysis, feature selection is a technique for selecting and filtering the most relevant features from text data to improve the performance of the sentiment analysis model. This method aims to improve computational efficiency, improve interpretability of results, reduce data dimensions, reduce overfitting, and improve predictability. The authors use datasets derived from the IMDb Review website, which has 10,000 review data with a ratio of 80% to 20% for training and test data for this study. In this study, sentimental analysis was conducted using the Support Vector Machine (SVM) algorithm, which used three feature selections: Mutual Information, N-Gram, and Chi-Square. This study aimed to look at the accuracy of comparisons of SVM sentimental analysis using feature selection and determine the most compelling feature selection method for SVM sentimental analysis. After that, each feature selection method is implemented with the same source and amount of data. The results of the applied tests showed that using the Support Vector Machine method for feature selection in spam text classification could improve classification performance. In this study, the classification of the Support Vector Machine algorithm without Feature Selection resulted in an accuracy of 85.35%. The results of the Mutual Information feature selection received an accuracy score of 90.00%, an accuracy on the N-Gram of 88.50%, and an accuracy on the Chi-Square of 88.65%. Thus, it can be concluded that testing and comparing the Support Vector Machine algorithm models with feature selection can improve classification performance. In this study, the Mutual Information feature selection score received the highest score with 90.00% accuracy.

Keywords : *Feature Selection, Sentiment Analysis, Support Vector Machine, IMDb Review, Mutual Information, N-Gram, Chi-Square*



ABSTRAK

Analisis sentimen adalah proses komputasional untuk mengekstrak, menemukan, atau menilai opini atau sentimen yang terkandung dalam teks, seperti ulasan, posting media sosial, artikel berita, atau jenis teks lainnya. Dalam analisis sentimen, seleksi fitur adalah teknik untuk memilih dan menyaring fitur-fitur yang paling relevan dari data teks untuk meningkatkan kinerja model analisis sentimen. Tujuan dari metode ini adalah untuk meningkatkan efisiensi komputasi, meningkatkan interpretabilitas hasil, mengurangi dimensi data, mengurangi overfitting, dan meningkatkan prediksi. Penulis menggunakan dataset yang berasal dari website IMDb Review, yang memiliki 10.000 data review dengan perbandingan 80% dan 20% untuk data latih dan data uji untuk penelitian ini. Pada penelitian ini, analisis sentiment dilakukan dengan algoritma *Support Vector Machine* (SVM) dan menggunakan tiga seleksi fitur, diantaranya *Mutual Information*, *N-Gram*, dan *Chi-Square*. Tujuan dari penelitian ini adalah untuk melihat komparasi accuracy analisis sentiment SVM memakai seleksi fitur dan tidak serta menentukan metode seleksi fitur yang paling efektif untuk analisis sentiment SVM. Setelah itu dilakukan komparasi setiap metode seleksi fitur yang di implementasikan dengan sumber dan jumlah data yang sama. Hasil pengujian yang diterapkan menunjukkan bahwa penggunaan metode *Support Vector Machine* untuk seleksi fitur dalam pengklasifikasian teks spam dapat meningkatkan kinerja klasifikasi. Dalam penelitian ini, hasil dari klasifikasi algoritma *Support Vector Machine* tanpa Seleksi Fitur menghasilkan *accuracy* 85,35%. Hasil dari seleksi fitur *Mutual Information* mendapatkan nilai *accuracy* sebesar 90,00%, *accuracy* pada *N-Gram* 88,50% dan *accuracy* pada *Chi-Square* 88,65%. Sehingga dapat disimpulkan bahwa hasil pengujian dan perbandingan model algoritma *Support Vector Machine* yang memiliki seleksi fitur dapat meningkatkan kinerja klasifikasi. Dalam penelitian ini, nilai seleksi fitur *Mutual Information* mendapatkan nilai tertinggi yaitu dengan *accuracy* 90,00%.

Kata kunci : Seleksi Fitur, Analisis Sentiment, *Support Vector Machine*, IMDb Review, *Mutual Information*, *N-Gram*, *Chi-Square*.



KATA PENGANTAR

Ucapan syukur *alhamdulillah* penulis panjatkan hanya kepada Allah *Subhanahu Wata'ala* yang telah banyak memberikan rahmat dan rahim berupa kesehatan, kekuatan, kesabaran dan juga ilmu pengetahuan yang dilimpahkan kepada manusia biasa yang banyak memiliki kekurangan. Sholawat serta salam "*Allahumma Sholli 'Alaa Sayyidina Muhammad Wa'ala Aali Sayyidina Muhammad*" juga penulis sampaikan kepada nabi Muhammad Shallallahu 'Alaihi Wasallam beserta keluarga dan sahabatnya.

Skripsi dengan judul "**Analisis Sentimen Komparasi Seleksi Fitur *Mutual Information, N-Gram, dan Chi-Square* Menggunakan Metode SVM Pada IMDb Review**" penulis hadirkan Skripsi ini sebagai salah satu syarat untuk menyelesaikan pendidikan Program Sarjana (S1) pada Program Studi Sistem Informasi, Fakultas Teknik Komputer dan Desain, Universitas Nusa Putra Sukabumi.

Penulis menyadari bahwa selama proses penyusunan skripsi ini membutuhkan kerja keras, usaha dan pengorbanan yang maksimal. Penulis juga menyadari bahwa Skripsi ini tidak akan terselesaikan tanpa do'a, dukungan, bantuan, serta bimbingan dan nasehat dari semua pihak. Oleh karena itu, penulis ingin menyampaikan penghargaan dan ucapan terimakasih yang sebesar-besarnya kepada:

1. Bapak Dr. Kurniawan, ST., M.Si, M.M selaku Rektor Universitas Nusa Putra Sukabumi.
2. Wakil Rektor I Bidang Akademik Universitas Nusa Putra Sukabumi Bapak Anggy Pradiftha Junfithrana, S.Pd, MT
3. Bapak Adhitia Erfina, S.T, M.Kom selaku Ketua Prodi Sistem Infromasi sekaligus Dosen Pembimbing I yang telah meluangkan waktu, tenaga dan pikiran untuk memberikan arahan, pengetahuan, bimbingan dan petunjuk yang sangat bermanfaat bagi penulis dalam menyusun skripsi ini
4. Bapak Dede Sukmawan, S.Kom, M.Kom selaku Dosen Pembimbing II yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan arahan, dan bimbingan bagi penulis dalam menyusun skripsi ini.



5. Dosen Penguji dalam sidang skripsi ini
6. Segenap Dosen Program Studi Sistem Informasi Nusa Putra Sukabumi yang telah memberikan pengalaman dan bekal ilmu pengetahuan yang bermanfaat selama penulis menempuh pendidikan di Program Studi Sistem Informasi Universitas Nusa Putra Sukabumi.
7. Kedua orang tua yang kami cintai, Ayah, Ibu, Adik, Kakak serta Keluarga yang senantiasa memberikan semangat, do'a, kasih harapan dan dorongan moral dan spritual kepada penulis dalam menyelesaikan tugas akhir skripsi.
8. Kepada rekan rekan Ismi Nurhadiani, Zesika Ayu, Rafli Radya, Trisyani Indah, Lusi, Santi Widiyawati, Windi Meliani, Muhammad Ilyas, Linda Wulandari, Wuli Anfarizah, Siska Tetra, Dimas Fitra, Muhammad Amara Dipta, Andri Sukmana, Arul dan Teman-teman KKN dan seluruh rekan rekan Mahasiswa angkatan 2020 Program Studi Sistem Informasi yang telah membantu, dan mendukung dalam pembuatan skripsi ini
9. Thanks to Me, My Self and I, Kepada kami bertiga, Sevira Chairunnisa, Fira Elfariani, Mirna Sari. Apresiasi sebesar besarnya karena telah bertanggung jawab untuk memulai dan menyelesaikan skripsi ini.
10. Semua orang baik dan berbagai pihak yang telah memberikan do'a, dukungan, motivasi dan saran yang tidak dapat disebutkan satu persatu dengan ikhlas dalam proses penyelesaian studi penulis.

Penulis menyadari bahwa Skripsi ini masih terdapat kekurangan dan jauh dari kesempurnaan, yang disebabkan oleh kurangnya ilmu dan wawasan yang dimiliki oleh penulis. Oleh karena itu, kritik dan saran yang membangun dari berbagai pihak akan lebih menyempurnakan Skripsi ini.

Sukabumi, 11 Juni 2024

Penulis

Penulis

Penulis

Sevira Chairunnisa N
20200050009

Fira Elfariani
20200050079

Mirna Sari Gusti R
20200050029



HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Nusa Putra, saya yang bertanda tangan dibawah ini :

Nama	: Sevira Chairunnisa Nandaresta	NIM : 20200050009
Nama	: Fira Elfariani	NIM : 20200050079
Nama	: Mirna Sari Gusti R	NIM : 20200050029
Program Studi	: Sistem Informasi	
Jenis Karya	: Skripsi	

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Nusa Putra **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty- Free Right*)** atas karya ilmiah saya yang berjudul:

“ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR *MUTUAL INFORMATION, N-GRAM* DAN *CHI-SQUARE* MENGGUNAKAN METODE SVM PADA *IMDB REVIEW*”

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Nusa Putra berhak menyimpan, mengalih media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya,

Dibuat di : Sukabumi

Pada tanggal : 11 Juni 2024

Yang Menyatakan

Penulis

Penulis

Penulis

Sevira Chairunnisa N
20200050009

Fira Elfariani
20200050079

Mirna Sari Gusti R
20200050029

DAFTAR ISI

HALAMAN COVER	i
HALAMAN JUDUL.....	i
PERNYATAAN PENULIS.....	ii
PENGESAHAN SKRIPSI	iii
ABSTRACT.....	iv
ABSTRAK.....	v
KATA PENGANTAR.....	vi
HALAMAN PERSETUJUAN PUBLIKASI	vii
DAFTAR ISI	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiv
DAFTAR LAMPIRAN.....	xv
BAB I PENDAHULUAN	1
10.1 Latar Belakang.....	1
10.2 Rumusan Masalah	2
10.3 Batasan Masalah	3
10.4 Tujuan Penelitian	3
10.5 Manfaat penelitian	3
10.6 Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	6
2.1 Penelitian Terkait.....	6
2.2 Landasan Teori	7
2.2.1 Kaggle.....	7
2.2.2 Analisis Sentimen	8
2.2.3 <i>Pre-Processing Text</i>	9
2.2.4 Algoritma <i>Support Vector Machine (SVM)</i>	10
2.2.5 Seleksi Fitur	12
2.2.6 <i>Confusion Matrix</i>	14
2.2.7 Bahasa Pemrograman <i>Python</i>	15



2.2.8 <i>Tools</i> Pendukung	15
2.3 Kerangka Pemikiran	16
BAB III METODOLOGI PENELITIAN.....	18
3.1 Tahapan Penelitian.....	18
3.2 Pengumpulan Data.....	19
3.3 <i>Pre-Processing Data</i>	19
3.3.1 <i>Cleansing</i>	20
3.3.2 <i>Case Folding</i>	21
3.3.3 <i>Tokenize</i>	21
3.3.4 <i>Remove Stopwords</i>	22
3.3.5 <i>Stemming</i>	23
3.4 Seleksi Fitur.....	23
3.5 Klasifikasi Data	24
3.6 Evaluasi Model.....	24
3.6.1 <i>Accuracy</i>	24
3.6.2 <i>Precision</i>	25
3.6.3 <i>Recall</i>	25
BAB IV HASIL DAN PEMBAHASAN	27
4.1 Pengumpulan Data.....	27
4.2 <i>Pre-Processing Data</i>	28
4.2.1 <i>Cleansing</i>	28
4.2.2 <i>Case Folding</i>	30
4.2.3 <i>Tokenize</i>	32
4.2.4 <i>Remove Stopwords</i>	33
4.2.5 <i>Stemming</i>	35
4.3 Klasifikasi dan Pengujian Model <i>Support Vector Machine</i>	36
4.3.1 Klasifikasi dan Pengujian Model <i>Support Vector Machine</i> tanpa Seleksi Fitur.....	36
4.3.2 Klasifikasi dan Pengujian Model <i>Support Vector Machine</i> dengan Seleksi Fitur	39





BAB V KESIMPULAN DAN SARAN.....	45
5.1 Kesimpulan	45
5.2 Saran	45
DAFTAR PUSTAKA.....	47
LAMPIRAN	50





DAFTAR GAMBAR

2.1	Gambar Kerangka Pemikiran.....	18
3.1	Gambar Alur Penelitian.....	18
3.2	Gambar Contoh Pengumpulan Data	19
3.3	Gambar Tahapan <i>Pre-processing</i>	20
3.4	Gambar Rumus Menghitung Nilai <i>Accuracy</i>	24
3.5	Gambar Rumus Menghitung Nilai <i>Precision</i>	25
3.6	Gambar Rumus Menghitung Nilai <i>Recall</i>	25
4.1	Gambar Website Kaggle	27
4.2	Gambar Kode Pemrograman <i>Cleansing</i>	29
4.3	Gambar Kode Pemrograman <i>Case Folding</i>	31
4.4	Gambar Kode Pemrograman <i>Tokenize</i>	32
4.5	Gambar Kode Pemrograman <i>Remove Stopwords</i>	34
4.6	Gambar Kode Pemrograman <i>Stemming</i>	35
4.7	Gambar Kode Pemrograman <i>Library Process</i>	37
4.8	Gambar Kode Pemrograman Klasifikasi tanpa Seleksi Fitur	37
4.9	Gambar Kode Pemrograman Klasifikasi tanpa Seleksi Fitur	38
4.10	Gambar Kode Pemrograman <i>Confusion Matrix</i> tanpa Seleksi Fitur	38
4.11	Gambar Hasil <i>Confusion Matrix</i> tanpa Seleksi Fitur	38
4.12	Gambar Hasil <i>Classification Report Support Vector Machine</i> tanpa Seleksi Fitur.....	39
4.13	Gambar Kode Pemrograman Klasifikasi memakai Seleksi Fitur <i>Mutual Information, N-Gram</i> dan <i>Chi-Square</i>	40
4.14	Gambar Kode Pemrograman Klasifikasi memakai Seleksi Fitur <i>Mutual Information</i>	40
4.15	Gambar Kode Pemrograman Klasifikasi memakai Seleksi Fitur <i>N-Gram</i>	40
4.16	Gambar Kode Pemrograman Klasifikasi memakai Seleksi Fitur <i>Chi-Square</i>	40



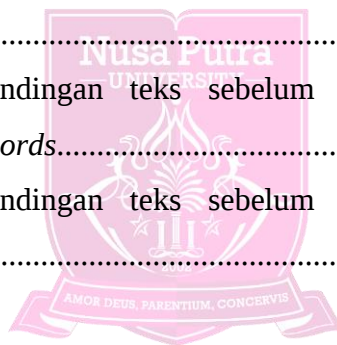
4.17	Gambar Kode Pemrograman <i>Confusion Matrix</i> pada Seleksi Fitur <i>Mutual Information</i> , <i>N-Gram</i> dan <i>Chi-Square</i>	41
4.18	Gambar Hasil Evaluasi <i>Confusion Matrix</i> pada Seleksi Fitur <i>Mutual Information</i>	41
4.19	Gambar Hasil Evaluasi <i>Confusion Matrix</i> pada Seleksi Fitur <i>N-Gram</i>	42
4.20	Gambar Hasil Evaluasi <i>Confusion Matrix</i> pada Seleksi Fitur <i>Chi-Square</i>	42
4.21	Gambar Hasil <i>Classification Report</i> menggunakan Seleksi Fitur <i>Mutual Information</i> , <i>N-Gram</i> dan <i>Chi-Square</i>	43





DAFTAR TABEL

3.1	Tabel Contoh Hasil <i>Cleansing</i>	20
3.2	Tabel Contoh Hasil <i>Case Folding</i>	21
3.3	Tabel Contoh Hasil <i>Tokenize</i>	22
3.4	Tabel Contoh Hasil <i>Remove Stopwords</i>	22
3.5	Tabel Contoh Hasil <i>Stemming</i>	23
4.1	Tabel Contoh Pengumpulan Data	28
4.2	Tabel Hasil perbandingan teks sebelum dan sesudah dilakukan proses <i>Cleansing</i>	30
4.3	Tabel Hasil perbandingan teks sebelum dan sesudah dilakukan proses <i>Case Folding</i>	31
4.4	Tabel Hasil perbandingan teks sebelum dan sesudah dilakukan proses <i>Tokenize</i>	33
4.5	Tabel Hasil perbandingan teks sebelum dan sesudah dilakukan proses <i>Remove Stopwords</i>	34
4.6	Tabel Hasil perbandingan teks sebelum dan sesudah dilakukan proses <i>Stemming</i>	36





DAFTAR LAMPIRAN

Biodata Mahasiswa.....	51
------------------------	----





BAB I

PENDAHULUAN

1.1 Latar Belakang

Data yang dihasilkan secara terus menerus melalui *platform* seperti media sosial, situs web, dan berbagai *platform* lainnya semakin melimpah di era teknologi modern. Dengan menggunakan data ini, kita dapat memperoleh pemahaman yang lebih baik tentang berbagai aspek kehidupan. Sentimen pengguna, atau pemikiran mereka tentang suatu objek atau entitas tertentu, adalah salah satu jenis informasi yang dapat diekstraksi.

Analisis sentimen adalah teknik yang digunakan untuk mengekstraksi dan menganalisis sentimen dari teks yang dibuat oleh pengguna. Tujuan utama dari analisis sentimen adalah untuk memahami sikap dan perasaan yang terkandung dalam teks, baik itu positif, negatif, atau netral. Dengan demikian, analisis sentimen memungkinkan perusahaan dan organisasi untuk mendapatkan wawasan yang bermanfaat tentang bagaimana pengguna merespon barang, layanan, atau peristiwa tertentu.

Salah satu komponen paling signifikan yang dapat mempengaruhi tingkat akurasi klasifikasi adalah seleksi fitur dataset karena dimensi ruang yang lebih besar akan mengurangi akurasi klasifikasi [1]. Dua kategori metode seleksi fitur adalah *filter* dan *wrapper*. Metode *wrapper* melakukan interaksi dengan algoritma klasifikasi untuk menentukan kegunaan fitur, sehingga menghasilkan sub fitur dengan kinerja yang lebih baik, namun, metode filter sendiri memerlukan evaluasi confusion matrix untuk mengetahui kemampuan fitur yang membedakan setiap kelas, sehingga cenderung lebih lambat. *Frequency of Document*, *Mutual Information*, *Information Gain*, dan *Chi-Square* adalah beberapa contoh metode filter yang sering digunakan dalam proses klasifikasi teks [2].

Penelitian yang ditulis oleh Jessica Widyadhana Iskandar dkk. berjudul "Perbandingan *Naive Bayes*, SVM, dan K-NN untuk Analisis Sentimen Gadget Berbasis Aspek" pada tahun 2021 bertujuan untuk membandingkan metode klasifikasi *Naive Bayes* (NB), *Support Vector Machine* (SVM), dan *K-Nearest*

Neighbor (K-NN) untuk menentukan model mana yang paling cocok untuk melakukan analisis dan klasifikasi gadget yang terkait. Untuk tujuan penelitian ini, metode model *Cross-Industry Standard Process for Data Mining* (CRISP-DM) digunakan. Hasil penelitian menunjukkan bahwa model klasifikasi *Support Vector Machine* (SVM) memberikan hasil terbaik dalam melakukan analisis sentimen terhadap komentar YouTube tentang Samsung Galaxy Z Flip 3 dibandingkan dengan metode *Naïve Bayes* (NB) dan *K-Nearest Neighbor* (K-NN) [3].

Analisis sentimen pada dataset *review* IMDb sendiri sudah banyak dilakukan, karena data tersebut merupakan data terbuka yang sering digunakan sebagai data rujukan dalam penelitian berbasis analisis sentimen. Setelah kami membaca beberapa jurnal tentang seleksi fitur, dari beberapa jurnal tersebut rata-rata hanya menggunakan 2 seleksi fitur, dan sayangnya belum ada penelitian yang membandingkan 3 metode seleksi fitur yang paling efektif pada algoritma SVM. Pada penelitian ini, kami ingin mencari keterbaruan untuk menentukan metode seleksi fitur mana yang paling efektif untuk analisis sentimen SVM dan setiap metode seleksi fitur yang telah digunakan sebelumnya harus dibandingkan dengan metode seleksi fitur lain dengan sumber dan jumlah data yang sama.

Berdasarkan latar belakang tersebut, penulis melakukan suatu penelitian yang berjudul **“ANALISIS SENTIMEN KOMPARASI SELEKSI FITUR *MUTUAL INFORMATION*, *N-GRAM*, DAN *CHI-SQUARE* MENGGUNAKAN METODE SVM PADA IMDb Review”**

1.2 Rumusan Masalah

Meskipun SVM telah terbukti efektif dalam banyak kasus, pertanyaan yang muncul adalah sejauh mana fitur – fitur yang ada dalam SVM dapat bersaing antara satu dengan lainnya yang mungkin menawarkan keunggulan dalam hal kinerja, kecepatan, atau skalabilitas. Oleh karena itu, perumusan masalah utama dalam penelitian ini adalah:

1. Bagaimana komparasi kinerja metode Support Vector Machine menggunakan Seleksi Fitur dan tidak menggunakan Seleksi Fitur?
2. Bagaimana perbandingan kinerja SVM dalam melakukan analisis sentimen dengan menggunakan berbagai jenis seleksi fitur yang berbeda?

3. Apa pengaruh dari penggunaan seleksi fitur *Mutual Information*, *N-Gram*, dan *Chi Square* terhadap akurasi klasifikasi SVM dalam analisis sentimen?

1.3 Batasan Masalah

Batasan masalah ini bertujuan agar penelitian lebih terarah sesuai dengan rumusan masalah serta menjadi batasan kemampuan penulis. Berikut batasan masalah pada penelitian ini dapat dilihat dibawah ini :

1. Penelitian ini akan berfokus pada komparasi seleksi fitur Analisis Sentimen pada algoritma *Support Vector Machine*
2. Data yang diambil dalam penelitian sebanyak 10.000 data pada *website* Kaggle dengan data latih 80% dan data uji 20%
3. Pada penelitian ini kami menggunakan *tools* VS Code versi 1.90.0. File *ipynb* dijalankan dengan VS Code menggunakan *Jupyter* 6.5.0 dan kernel *Python* 3.11.9.
4. Hasil dari pengklasifikasian di dalam analisis ini berupa sentimen positif dan negatif

1.4 Tujuan Penelitian

Tujuan dari penulisan penelitian ini adalah:

1. Untuk mengetahui komparasi nilai akurasi dan kinerja SVM dengan menggunakan Seleksi Fitur dan tidak menggunakan Seleksi Fitur
2. Untuk mengetahui komparasi kinerja SVM dalam melakukan analisis sentimen dengan menggunakan berbagai jenis Seleksi Fitur yang berbeda.
3. Untuk mengetahui bagaimana pengaruh dari penggunaan Seleksi Fitur *Mutual Information*, *N-Gram*, dan *Chi Square* terhadap akurasi klasifikasi SVM dalam analisis sentiment

1.5 Manfaat penelitian

Penelitian mengenai perbandingan seleksi fitur dalam metode Support Vector Machine (SVM). Untuk analisis sentimen memiliki beberapa manfaat yang signifikan, di antaranya:

1. Penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang peran dan pengaruh dari berbagai seleksi fitur yang digunakan dalam SVM terhadap kinerja analisis sentimen.
2. Penelitian ini dapat membantu mengidentifikasi fitur-fitur yang paling berpengaruh dan kombinasi fitur yang optimal dapat membantu meningkatkan akurasi dan efisiensi klasifikasi sentimen.
3. Penelitian ini dapat membantu dalam menghemat sumber daya komputasi yang diperlukan untuk melatih dan menguji model SVM. Fitur-fitur yang kurang relevan atau tidak berpengaruh dapat dihindari, sehingga mengurangi waktu dan biaya komputasi yang diperlukan.
4. Temuan dari penelitian ini dapat menjadi landasan bagi penelitian lanjutan, hal ini dapat mendorong terciptanya inovasi baru dalam penggunaan SVM dan fitur-fitur yang relevan dalam analisis sentimen.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran umum terhadap penelitian yang dilakukan, sistematika penulisan Tugas Akhir ini terdiri dari 5 (lima) bab, yaitu:

BAB I : Pendahuluan

Meliputi uraian mengenai Topik, Latar Belakang, Rumusan Masalah, Tujuan Penelitian, Batasan Masalah, Metodologi penelitian, Sistematika Penelitian, dan Ruang Lingkup Penelitian.

BAB II : Tinjauan Pustaka

Berisi Tentang Peneliti Terkait, Landasan Teori dan Kerangka Pemikir.

BAB III : Metodologi Penelitian

Isi dari bab ini adalah pengumpulan data dan metode yang digunakan saat penelitian.

BAB IV Hasil Penelitian

Menjelaskan dari Hasil dan Pembahasan Penelitian Serta

: Implementasi Sistem.

BAB V Penutup

: Berisi tentang kesimpulan dan saran.





BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian mengenai analisis sentimen yang dilakukan, peneliti menggunakan open data dari website Kaggle dengan total 10.000 data *review* dengan menggunakan perbandingan 80% dan 20% untuk data latih dan data uji. Algoritma *Support Vector Machine*, bahasa pemrograman *Python* dan tools pendukung *Visual Studio Code* digunakan pada penelitian ini. Hasil dari klasifikasi algoritma *Support Vector Machine* tanpa Seleksi Fitur dengan bahasa pemrograman *Python* menghasilkan 1.016 data sentiment positif, dan 984 data sentiment negatif dengan nilai *accuracy* 85,35%.

Hasil dari klasifikasi algoritma *Support Vector Machine* dengan implementasi seleksi fitur didapatkan 1.515 sentimen positif dan 485 sentimen negatif pada semua seleksi fitur. Hasil dari seleksi fitur *Mutual Information* mendapatkan nilai *accuracy* sebesar 90,00%, hasil nilai *accuracy* pada *N-Gram* 88,50% dan hasil nilai *accuracy* pada *Chi-Square* 88,65%. Berdasarkan hasil dari penelitian membuktikan seleksi fitur *Mutual Information* mendapatkan nilai tertinggi dibandingkan model fitur lainnya. Sehingga dapat disimpulkan bahwa hasil pengujian dan perbandingan klasifikasi algoritma *Support Vector Machine* pada seleksi fitur *Mutual Information* memiliki nilai *accuracy* paling tinggi dengan selisih tertinggi 4,65% begitupun dengan kedua model lainnya yang memiliki nilai seleksi yang lebih tinggi dibandingkan dengan klasifikasi sentimen *Support Vector Machine* yang tidak memiliki seleksi fitur, dan terbukti bahwa penggunaan seleksi fitur dapat meningkatkan kinerja klasifikasi dibandingkan dengan yang tidak memiliki seleksi fitur.

5.2 Saran

Penulis menyadari bahwa masih banyak kekurangan dan kelemahan dalam penelitian ini terutama dalam proses pengumpulan data, pengolahan data, *pre-processing text* dan khususnya pada proses Seleksi Fitur. Oleh karena itu peneliti selanjutnya yang ingin melakukan penelitian dengan topik serupa sebaiknya

memperhatikan lagi kualitas dan jumlah data, pengambilan data sebaiknya bukan dari opendata, lebih baik pengambilan data diambil dari Youtube, Google Play Store atau data yang diambil melalui proses *crawling* data karena akan meminimalisir plagiarisme saat hasil pre-processing ditampilkan dalam tabel di hasil dan pembahasan dan data lebih *fresh*, dan saran selanjutnya disarankan menggunakan metode selain *Support Vector Machine* (SVM) , dan Fitur Seleksi selain *Mutual Information*, *N-Gram*, dan *Chi-Square* untuk keterbaharuan penelitian selanjutnya.



DAFTAR PUSTAKA

- [1]. Tajul Arifin AMIK BSI Bogor Jl Merdeka No, Y., & Jawa Barat, B. (2016). KOMPARASI FITUR SELEKSI PADA ALGORITMA SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN REVIEW. INFORMATIKA, 3, 191–199.
<https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/868>
- [2]. ARUDA, S. A. Q. (n.d.). PERBANDINGAN METODE SELEKSI FITUR UNTUK KLASIFIKASI PERTANYAAN BERBAHASA INDONESIA MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM).
- [3]. Iskandar, J. W., & Nataliani, Y. (2021). Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(6), 1120–1126.
<https://doi.org/10.29207/resti.v5i6.3588>
- [4]. al Qodrin, S., Yusliani, N., & Syahrini, A. (n.d.). Bulan Oktober Tahun 2022, Hal. In Jurnal JUPITER (Vol. 14, Issue 2).
- [5]. Arifin, N., Enri, U., & Sulistiyowati, N. (n.d.). STRING (Satuan Tulisan Riset dan Inovasi Teknologi) PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM UNTUK TEXT CLASSIFICATION.
- [6]. Nasib, S. K., Pammus, F. I., Nurwan, & Nashar, L. O. (2023). COMPARISON OF FEATURE SELECTION BASED ON COMPUTATION TIME AND CLASSIFICATION ACCURACY USING SUPPORT VECTOR MACHINE. Indonesian Journal of Applied Research (IJAR), 4(1), 63–74. <https://doi.org/10.30997/ijar.v4i1.252>
- [7]. Lestari, S., Mupaat, M., Erfina, A., Studi, P., Informasi, S., Sains, F., & Teknologi, D. (n.d.). Analisis Sentimen Masyarakat Indonesia terhadap Pemindahan Ibu Kota Negara Indonesia pada Twitter. 8(1), 13–22.
- [8]. Fahlevvi, M. R. (2022). Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi dan Dokumentasi Kementerian Dalam Negeri

- Republik Indonesia di Google Playstore Menggunakan Metode Support Vector Machine. *Jurnal Teknologi Dan Komunikasi Pemerintahan*, 4(1), 1–13.
- [9]. Fauzan, M. A. (n.d.). PENERAPAN SENTIMEN LEKSIKON INDONESIA PADA ANALISIS SENTIMEN MENGENAI OPINI MASYARAKAT DI TWITTER TERHADAP KEBIJAKAN PPKM DARURAT MENGGUNAKAN ALGORITMA MAXIMUM ENTROPY SKRIPSI Oleh.
- [10]. Alfredo Tpoy, G., Muliantara, A., Raya Kampus Udayana, J., Jimbaran, B., Selatan, K., & Indonesia, B. (n.d.). Klasifikasi Teks Spam dengan Algoritma Support Vector Machine dan Chi-Square. In *JNATIA* (Vol. 1, Issue 4).
- [11]. Mubarak, H., Ernawati, I., & Chamidah, N. (2022). Optimasi Algoritma Support Vector Machine Menggunakan Seleksi Fitur Particle Swarm Optimization Pada Analisis Sentimen Terhadap Kebijakan PPKM.
- [12]. Alfredo Tpoy, G., Muliantara, A., Raya Kampus Udayana, J., Jimbaran, B., Selatan, K., & Indonesia, B. (n.d.). Klasifikasi Teks Spam dengan Algoritma Support Vector Machine dan Chi-Square. In *JNATIA* (Vol. 1, Issue 4).
- [13]. Suri, D. J., & Purnamasari, K. K. (n.d.). Perbandingan Seleksi Fitur Untuk Klasifikasi Sentimen SVM Pada Twitter.
- [14]. Kurnia, R. P., & Atma, Y. A. (n.d.). ANALISIS REKOMENDASI FILM DARI DATA IMDB MENGGUNAKAN PYTHON ANALYSIS OF FILM RECOMMENDATIONS FROM IMDB DATA USING PYTHON. 3(2), 23–28.
- [15]. Ramanda Hasibuan, M. (2019). Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN) (Vol. 3, Issue 11). <http://j-ptiik.ub.ac.id>
- [16]. Radit, A. P., Agnes, D., & Polina, M. (n.d.). Analisis Sentimen Terhadap Penggunaan Marketplace di Indonesia Menggunakan Metode Support Vector Machine dengan Seleksi Fitur Chi Square. <https://github.com/onpilot/sentimen-bahasa>

- [17]. Priya, S., Reddy, V. D., & Balaji, V. (2022). Using AI to detect and classify malicious domain names. *International Journal of Health Sciences*, 9538–9547. <https://doi.org/10.53730/ijhs.v6ns1.7198>
- [18]. Hartanto, P., Saepudin, S., Rahmalenia, D., Maulinda, D., Fauziah, T., & Ocktafiani, N. (n.d.). Indonesia's Economic Recovery Post Covid-19 Pandemic Sentiment Analysis 1 st Adhitia Erfina 2 nd Lakoa Fitina 3 rd.
- [19]. Suri, D. J., & Purnamasari, K. K. (n.d.). Perbandingan Seleksi Fitur Untuk Klasifikasi Sentimen SVM Pada Twitter.
- [20]. Natasuwarna, A. P. (2020). Seleksi Fitur Support Vector Machine pada Analisis Sentimen Keberlanjutan Pembelajaran Daring. *Techno.Com*, 19(4), 437–448. <https://doi.org/10.33633/tc.v19i4.4044>
- [21]. Erfina, A., Maulana, R., Agustian, R., & Nurlatipah, T. R. (n.d.). SENTIMENT ANALYSIS OF INDONESIAN TWITTER ON THE GOVERNMENT IN HANDLING THE COVID19 VIRUS. 2–7.
- [22]. Infomation, S. (n.d.). Indonesian Twitter Sentiment Analysis Application on The Covid19 Vaccine Using PHP and Naive Bayes Classifier.
- [23]. Erfina, A., Tamanin, A. S., Listianawati, Sembiring, F., Saepudin, S., Melawati, & Lesmana, C. T. (2020). New Approach of Sarcasm Detection in Indonesian Marketplace Product Review. 6th International Conference on Computing, Engineering, and Design, ICCED 2020. <https://doi.org/10.1109/ICCED51276.2020.9415787>
- [24]. Erfina, A., Kamila, S. N. M., Nurushalihah, N., Tsaqila, S. A., Sukmawan, D., & Saepudin, S. (2023). Big Asian Countries Economic Recovery Post Covid-19 Pandemic Sentiment Analysis. 2023 IEEE 9th International Conference on Computing, Engineering and Design, ICCED 2023. <https://doi.org/10.1109/ICCED60214.2023.10425216>
- [25]. Mubarok, H., Ernawati, I., & Chamidah, N. (2022). Optimasi Algoritma Support Vector Machine Menggunakan Seleksi Fitur Particle Swarm



